0-315-01034-7

National Library of Canada
Canadian Theses Division
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Division des thèses canadiennes

48892

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

- Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur
David E. Blackmore

Date of Birth — Date de naissance
Dec 2, 1946

Country of Birth — Lieu de naissance
Canada

Permanent Address — Résidence fixe
11720 - 48 Ave
Edmonton, Alberta
T6H 0E6

Title of Thesis — Titre de la thèse
A Latent Trait Study of Item Bias and Achievement Differences

University — Université
University of Alberta

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée
Ph.D.

Year this degree conferred — Année d'obtention de ce grade
Fall, 1980

Name of Supervisor — Nom du directeur de thèse
T. O. Maguire

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date
9/17/80

Signature

NL-91 (4/77)

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

Ottawa, Canada
K1A 0N4

NL-339 (Rev. 8/80)

THE UNIVERSITY OF ALBERTA

A LATENT TRAIT STUDY OF ITEM BIAS AND ACHIEVEMENT DIFFERENCES

by

© David E. Blackmore

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF Doctor of Philosophy

Educational Psychology

EDMONTON, ALBERTA
Fall, 1980

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled A LATENT TRAIT STUDY OF ITEM BIAS AND ACHIEVEMENT DIFFERENCES submitted by David E. Blackmore in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

T. Maguire

Supervisor

V.R. Nyberg

External Examiner

Date....May 30, 1980....

This thesis is dedicated to
Marva, Eric, and Beth-Ann

ABSTRACT

The test results that are obtained from a test-retest design, whether or not the same sample was employed in the two test administrations, are often compared in order to determine an achievement difference score. But, can the researcher be assured that the same latent trait was being measured each time that the test was administered?

The purpose of this thesis was to demonstrate an application of latent trait theory which enables the researcher to assess whether or not those tests that are employed in a test-retest design are measuring the same latent trait each time that the test was administered. The methodology was devised from the premise that a test item should assess achievement in a similar manner on each test administration, across all levels of student ability. This assessment was made by computing a probable item difference score between the two groups of examinees for each test item by first computing the equated item characteristic curves and then multiplying the differences (between groups) of the probability of success for a given student ability level by the proportion of total scores that were found at that ability level; then summing across all ability levels. Those items that were found to have a probable item difference score of greater than 0.05 were considered to be aberrant items that should not be included in the final mean achievement difference score; i.e., the examinees having the

same ability level in each study group that the test was administered to, did not have the same probability of correctly answering the aberrant item.

In order to illustrate this application of latent trait theory, a portion of the data that were collected for the Edmonton Grade III Achievement Study (Clarke, Nyberg, & Worth, 1977) was employed. This study administered a battery of tests to the Edmonton grade III public school students in 1956 and again to the Edmonton grade III public school students in 1977 in order to determine achievement differences across 21 years. The results indicate that there were items within this test battery that were not measuring the same latent trait each time that they were administered. When the data from the aberrant items were removed, the 1977 students appeared to have higher abilities, as measured by the selected tests, than did the 1956 students.

ACKNOWLEDGEMENTS

The author wishes to express sincere thanks to Dr. T. O. Maguire, Dr. W. N. Runquist, Dr. V. Nyberg, Dr. G. Kysela, and Dr. W. Muir, who served on the thesis committee and made the completion of this project possible.

An acknowledgement of gratitude is also extended to Dr. S. C. T. Clarke, Dr. V. Nyberg, and Dr. W. Worth for the opportunity to have been part of the Grade III Achievement Study and to have been able to have access to the data collected from that study.

A special note of gratitude must be extended to Dr. T. O. Maguire for his support and understanding that went beyond the call of duty. A special note of thanks must also be given to Dr. S. Hunka, Lana Ray Southorn, and the many devoted staff members of the Division of Educational Research for their continued support, and for providing the opportunity to learn so many new, interesting, and necessary things.

In addition, I would like to thank Emma Collins for the help and friendship that she has extended to me during my residency at the University of Alberta and the Department of Educational Psychology for providing the computing funds that were necessary in order for me to complete this project.

To my parents, wife, and family, to whom this thesis is dedicated, I give special thanks, for their much needed love, encouragement, suggestions, and understanding.

Table of Contents

Chapter Page

I. INTRODUCTION 1

Purpose 6

II. REVIEW OF RELATED LITERATURE 9

A. Latent Trait Theory 9

Dimensionality 10

Local Independence 11

Item Characteristic Curve 12

Different Types of Latent Trait Models 13

Methods of Obtaining the Item Characteristic Curve Parameters 17

Applications of Latent Trait Theory 20

B. Edmonton Grade III Achievement Data 24

Data Collection Procedures 28

C. Chapter Summary 29

III. METHOD 31

Data Analysis 31

IV. RESULTS 40

Raven Progressive Matrices Test 42

Gates Paragraph Reading 43

Gates Word Recognition 50

California Mental Maturity 57

V. DISCUSSION AND CONCLUSIONS 82

A. Introduction 82

B. Discussion of Individual Test Results 85

Raven Progressive Matrices Test 85

Gates Paragraph Reading Test 94

Gates Word Recognition Test 100

California Mental Maturity Test 108

C. Classical Test Theory Implications versus Latent Trait Theory Implications 118

D. Conclusions 124

Figures 130

References 279

APPENDIX A 284

LIST OF TABLES

Table Description Page

1 Study Item Number Assignment for Raven's Progressive Matrices Test 44

2 Original Item Parameters for Ravens's Progressive Matrices Test 45

3 Rescaled Item Parameters for the Raven Progressive Matrices Test 46

4 The Number of Students Located at Each Level of Ability for the Raven Progressive Matrices Test 47

5 Aberrant Items at Specified levels of Difference: Raven's Progressive Matrices Test . 48

6 Mean Student Abilities for Raven's Progressive Matrices Test 49

7 Study Item Number Assignment for Gates Paragragh Reading 51

8 Original Item Parameters for Gates Paragraph Reading 52

9 Rescaled Item Parameters for Gates Paragraph Reading 53

10 The Number of Students Located at Each Specified Level of Ability for the Gates Paragraph Reading 54

11 Aberrant Items at Specified Levels of Difference: Gates Paragraph Reading 55

LIST OF TABLES continued

Table	Description	Page
12	Mean Student Abilities for the Gates Paragraph Reading Test	56
13	Study Item Number Assignment for Gates Word Recognition	58
14	Original Item Parameters for Gates Word Recognition	59
15	Rescaled Item Parameters for Gates Word Recognition	61
16	The Number of Students Located at Each Level of Ability for the Gates Word Recognition Test . .	63
17	Aberrant Items at Specified Levels of Difference: Gates Word Recognition Test	64
18	Mean Student Abilities for the Gates Word Recognition Test	66
19	Study Item Number Assignment for California Mental maturity	68
20	Original Item Parameters for California Mental Maturity	71
21	Rescaled Item Parameters for California Mental Maturity	74
22	The Number of Students Located at Each Level of Ability for the California Mental Maturity Test	77
23	Aberrant Items at Specified Levels of Difference: California Mental Maturity	78

LIST OF TABLES continued

Table	Description	Page
24	Mean Student Abilities for the California Mental Maturity Test	81
25	Location of Item Bias by Ability Level for Raven's Progressive Matrices Test	87
26	The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for Raven's Progressive Matrices	93
27	Location of Item Bias by Ability Level for Gates Paragraph Reading Test	95
28	The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for Gates Paragraph Reading	99
29	Location of Item Bias by Ability Level for Gates Word Recognition	102
30	The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for Gates Word Recognition	107
31	Location of Item Bias by Ability Level for the California Mental Maturity Test	111
32	The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for the California Mental Maturity Test	117
33	Classical Test Theory Item Statistics for Raven's Progressive Matrices Test	119

LIST OF FIGURES

Figure	Description	Page
1	ICC'S FOR ITEM: b ≤ 4.00	131
2	ICC'S FOR ITEM NUMBER 7: RAVEN	132
3	ICC'S FOR ITEM NUMBER 8: RAVEN	133
4	ICC'S FOR ITEM NUMBER 9: RAVEN	134
5	ICC'S FOR ITEM NUMBER 10: RAVEN	135
6	ICC'S FOR ITEM NUMBER 11: RAVEN	136
7	ICC'S FOR ITEM NUMBER 12: RAVEN	137
8	ICC'S FOR ITEM NUMBER 16: RAVEN	138
9	ICC'S FOR ITEM NUMBER 17: RAVEN	139
10	ICC'S FOR ITEM NUMBER 18: RAVEN	140
11	ICC'S FOR ITEM NUMBER 19: RAVEN	141
12	ICC'S FOR ITEM NUMBER 20: RAVEN	142
13	ICC'S FOR ITEM NUMBER 21: RAVEN	143
14	ICC'S FOR ITEM NUMBER 22: RAVEN	144
15	ICC'S FOR ITEM NUMBER 23: RAVEN	145
16	ICC'S FOR ITEM NUMBER 24: RAVEN	146
17	ICC'S FOR ITEM NUMBER 27: RAVEN	147
18	ICC'S FOR ITEM NUMBER 28: RAVEN	148
19	ICC'S FOR ITEM NUMBER 29: RAVEN	149
20	ICC'S FOR ITEM NUMBER 30: RAVEN	150
21	ICC'S FOR ITEM NUMBER 31: RAVEN	151
22	ICC'S FOR ITEM NUMBER 32: RAVEN	152
23	ICC'S FOR ITEM NUMBER 33: RAVEN	153
24	ICC'S FOR ITEM NUMBER 34: RAVEN	154

LIST OF FIGURES continued

Figure Description Page

25 ICC'S FOR ITEM NUMBER 35: RAVEN 155

26 ICC'S FOR ITEM NUMBER 36: RAVEN 156

27 ICC'S FOR ITEM NUMBER 7: GPR 157

28 ICC'S FOR ITEM NUMBER 8: GPR 158

29 ICC'S FOR ITEM NUMBER 10: GPR 159

30 ICC'S FOR ITEM NUMBER 12: GPR 160

31 ICC'S FOR ITEM NUMBER 13: GPR 161

32 ICC'S FOR ITEM NUMBER 14: GPR 162

33 ICC'S FOR ITEM NUMBER 15: GPR 163

34 ICC'S FOR ITEM NUMBER 16: GPR 164

35 ICC'S FOR ITEM NUMBER 17: GPR 165

36 ICC'S FOR ITEM NUMBER 19: GPR 166

37 ICC'S FOR ITEM NUMBER 20: GPR 167

38 ICC'S FOR ITEM NUMBER 21: GPR 168

39 ICC'S FOR ITEM NUMBER 22: GPR 169

40 ICC'S FOR ITEM NUMBER 23: GPR 160

41 ICC'S FOR ITEM NUMBER 24: GPR 171

42 ICC'S FOR ITEM NUMBER 25: GPR 172

43 ICC'S FOR ITEM NUMBER 26: GPR 173

44 ICC'S FOR ITEM NUMBER 27: GPR 174

45 ICC'S FOR ITEM NUMBER 28: GPR 175

46 ICC'S FOR ITEM NUMBER 29: GPR 176

47 ICC'S FOR ITEM NUMBER 30: GPR 177

48 ICC'S FOR ITEM NUMBER 31: GPR 178

LIST OF FIGURES continued

Figure Description Page

49 ICC'S FOR ITEM NUMBER 32: GPR 179

50 ICC'S FOR ITEM NUMBER 6: GWR 180

51 ICC'S FOR ITEM NUMBER 10: GWR 181

52 ICC'S FOR ITEM NUMBER 11: GWR 182

53 ICC'S FOR ITEM NUMBER 12: GWR 183

54 ICC'S FOR ITEM NUMBER 13: GWR 184

55 ICC'S FOR ITEM NUMBER 14: GWR 185

56 ICC'S FOR ITEM NUMBER 15: GWR 186

57 ICC'S FOR ITEM NUMBER 16: GWR 187

58 ICC'S FOR ITEM NUMBER 17: GWR 188

59 ICC'S FOR ITEM NUMBER 18: GWR 189

60 ICC'S FOR ITEM NUMBER 19: GWR 190

61 ICC'S FOR ITEM NUMBER 20: GWR 191

62 ICC'S FOR ITEM NUMBER 21: GWR 192

63 ICC'S FOR ITEM NUMBER 22: GWR 193

64 ICC'S FOR ITEM NUMBER 23: GWR 194

65 ICC'S FOR ITEM NUMBER 24: GWR 195

66 ICC'S FOR ITEM NUMBER 25: GWR 196

67 ICC'S FOR ITEM NUMBER 26: GWR 197

68 ICC'S FOR ITEM NUMBER 27: GWR 198

69 ICC'S FOR ITEM NUMBER 28: GWR 199

70 ICC'S FOR ITEM NUMBER 29: GWR 200

71 ICC'S FOR ITEM NUMBER 30: GWR 201

72 ICC'S FOR ITEM NUMBER 31: GWR 202

LIST OF FIGURES continued

Figure Description Page

73 ICC'S FOR ITEM NUMBER 32: GWR 203

74 ICC'S FOR ITEM NUMBER 33: GWR 204

75 ICC'S FOR ITEM NUMBER 34: GWR 205

76 ICC'S FOR ITEM NUMBER 35: GWR 206

77 ICC'S FOR ITEM NUMBER 36: GWR 207

78 ICC'S FOR ITEM NUMBER 37: GWR 208

79 ICC'S FOR ITEM NUMBER 38: GWR 209

80 ICC'S FOR ITEM NUMBER 39: GWR 210

81 ICC'S FOR ITEM NUMBER 40: GWR 211

82 ICC'S FOR ITEM NUMBER 41: GWR 212

83 ICC'S FOR ITEM NUMBER 42: GWR 213

84 ICC'S FOR ITEM NUMBER 43: GWR 214

85 ICC'S FOR ITEM NUMBER 44: GWR 215

86 ICC'S FOR ITEM NUMBER 45: GWR 216

87 ICC'S FOR ITEM NUMBER 46: GWR 217

88 ICC'S FOR ITEM NUMBER 48: GWR 218

89 ICC'S FOR ITEM NUMBER 1: CMM 219

90 ICC'S FOR ITEM NUMBER 2: CMM 220

91 ICC'S FOR ITEM NUMBER 3: CMM 221

92 ICC'S FOR ITEM NUMBER 4: CMM 222

93 ICC'S FOR ITEM NUMBER 5: CMM 223

94 ICC'S FOR ITEM NUMBER 7: CMM 224

95 ICC'S FOR ITEM NUMBER 8: CMM 225

96 ICC'S FOR ITEM NUMBER 9: CMM 226

LIST OF FIGURES continued

Figure	Description	Page
97	ICC'S FOR ITEM NUMBER 10: CMM	227
98	ICC'S FOR ITEM NUMBER 11: CMM	228
99	ICC'S FOR ITEM NUMBER 12: CMM	229
100	ICC'S FOR ITEM NUMBER 13: CMM	230
101	ICC'S FOR ITEM NUMBER 14: CMM	231
102	ICC'S FOR ITEM NUMBER 15: CMM	232
103	ICC'S FOR ITEM NUMBER 16: CMM	233
104	ICC'S FOR ITEM NUMBER 17: CMM	234
105	ICC'S FOR ITEM NUMBER 18: CMM	235
106	ICC'S FOR ITEM NUMBER 19: CMM	236
107	ICC'S FOR ITEM NUMBER 20: CMM	237
108	ICC'S FOR ITEM NUMBER 21: CMM	238
109	ICC'S FOR ITEM NUMBER 22: CMM	239
110	ICC'S FOR ITEM NUMBER 23: CMM	240
111	ICC'S FOR ITEM NUMBER 24: CMM	241
112	ICC'S FOR ITEM NUMBER 25: CMM	242
113	ICC'S FOR ITEM NUMBER 27: CMM	243
114	ICC'S FOR ITEM NUMBER 28: CMM	244
115	ICC'S FOR ITEM NUMBER 29: CMM	245
116	ICC'S FOR ITEM NUMBER 30: CMM	246
117	ICC'S FOR ITEM NUMBER 31: CMM	247
118	ICC'S FOR ITEM NUMBER 32: CMM	248
119	ICC'S FOR ITEM NUMBER 33: CMM	249
120	ICC'S FOR ITEM NUMBER 36: CMM	250

LIST OF FIGURES continued

Figure Description Page

121 ICC'S FOR ITEM NUMBER 39: CMM 251

122 ICC'S FOR ITEM NUMBER 40: CMM 252

123 ICC'S FOR ITEM NUMBER 41: CMM 253

124 ICC'S FOR ITEM NUMBER 42: CMM 254

125 ICC'S FOR ITEM NUMBER 45: CMM 255

126 ICC'S FOR ITEM NUMBER 46: CMM 256

127 ICC'S FOR ITEM NUMBER 53: CMM 257

128 ICC'S FOR ITEM NUMBER 54: CMM 258

129 ICC'S FOR ITEM NUMBER 55: CMM 259

130 ICC'S FOR ITEM NUMBER 56: CMM 260

131 ICC'S FOR ITEM NUMBER 57: CMM 261

132 ICC'S FOR ITEM NUMBER 59: CMM 262

133 ICC'S FOR ITEM NUMBER 62: CMM 263

134 ICC'S FOR ITEM NUMBER 63: CMM 264

135 ICC'S FOR ITEM NUMBER 64: CMM 265

136 ICC'S FOR ITEM NUMBER 65: CMM 266

137 ICC'S FOR ITEM NUMBER 69: CMM 267

138 ICC'S FOR ITEM NUMBER 70: CMM 268

139 ICC'S FOR ITEM NUMBER 80: CMM 269

140 ICC'S FOR ITEM NUMBER 87: CMM 270

141 ICC'S FOR ITEM NUMBER 89: CMM 271

142 ICC'S FOR ITEM NUMBER 90: CMM 272

143 ICC'S FOR ITEM NUMBER 93: CMM 273

144 ICC'S FOR ITEM NUMBER 94: CMM 274

LIST OF FIGURES continued

Figure	Description	Page
145	ICC'S FOR ITEM NUMBER 95: CMM	275
146	ICC'S FOR ITEM NUMBER 96: CMM	276
147	ICC'S FOR ITEM NUMBER 97: CMM	277
148	ICC'S FOR ITEM: b ≥ 4.00	278

I. INTRODUCTION

There are many situations that arise in education, industry, and general research that necessitate the comparison and evaluation of two or more test scores obtained at different times. Tyler (1942) points out that presumably

> each of these scores or verbal summaries could be compared with scores or verbal summaries previously obtained; by this comparison some estimate of the degree of change or growth of students could be made.

While such comparisons can be made, a question that arises is whether or not the evaluator can be assured that the test(s) are measuring the same attribute or trait in each testing situation. Common sense would dictate that if two or more measurements are to be compared, then it is imperative that the test(s) measure the same underlying construct or trait, and that the results be reported in the same metric. However, history has demonstrated that this intuitive approach is easier to state than substantiate (e.g., Lumsden, 1976). This may be accounted for by the very nature of the indirect measurements required of psychological researchers. Traditionally, the physical scientist had few problems when making comparisons between test results of the same metric because he obtained his test results via direct measurement; e.g., the measurement of the physical size of an object. Thus, the psychometrician is confronted with the uncertainties inherent in the use of psychological measurements. These measurements are not only indirect, but they are often based upon deductions from hypothetical constructs of what might constitute a particular human trait or measurable attribute. For example, because intelligence is not a tangible substance, a psychologist measures it indirectly; via an instrument that is based on a belief about what intelligence is as an hypothetical construct; and how it is manifested in some observable fashion.

Educational or psychological testing, then, usually involves the translation of some unobservable human characteristics defined as traits, constructs, or abilities.[1] Because these traits are unobservable, they are often referred to as being latent. A person's score on an achievement test can then be described by his/her standing on a particular latent trait (e.g., mathematical ability, verbal ability, or motivational level).

In simple physical measurement, we believe that a ruler measures length with the same validity regardless of whether or not a desk or a field is being measured. In educational measurement, even if we are convinced that the same trait is being measured each time a given test is administered, we do not necessarily know that the obtained measure will be invariant over samples; that is, it might be measuring achievement in different ways for different samples.

In education, researchers may want to assess whether or not students of today know as much about some subject as other students before them knew (e.g., Muir, 1961; Hedges, 1977). If 20 minutes, 20 days, or 20 years have passed since the original students were measured on some ability, can the researcher be assured that the same test is measuring the same latent trait(s) that was measured before? Or, put another way, is the latent trait being measured in the same way?

---

[1] These terms are usually used interchangeably because there are no universally accepted definitions.

If the same underlying trait is being assessed each time the test is administered, then one would hope that the items composing the test would contribute the same type of information to the overall test score. Thus, if an item is assessing achievement in different ways on different administrations, then the test may be favoring or biasing the test results on the basis of which sample was administered the test. As the total score is a function of the individual item scores; the total amount of test bias that exists is a function of the bias contributed by each item. If the test results obtained for two or more samples are going to be compared, then a test item should not be biased toward any sample on any one specfic test administration.

Classical test theory and latent trait theory have both been used to establish methods of determining item bias (e.g., Lord & Novick, 1968; Lord, 1952, 1976a). Classical test theory is based upon a model that advocates that every "observed" score is composed of a "true" score component and an "error" score component (O = T + E) (Thurstone, 1931;

Gulliksen, 1950). While this model (O = T + E) makes intuitive sense, there are some problems that must be contended with; such as, the true score is not known and can not ever be known; and there is no assumption made about the frequency distribution of true scores in the sample to which it is applied. Thus, the measures used to assess item bias are dependent upon the group performance from which they were derived. Or, in the words of Hambleton, Swaminathan, Cook, Eignor, and Gifford (1977),

> We create a situation of bias and then try to use the mechanism that created the problem in the first place to investigate it [item bias].

Two measures are used to determine if a test item is performing in a similar manner each time that it is administered. The first measure is item difficulty which is defined as the porportion of examinees who answered an item correctly. The second measure is item discrimination which is a measure of how well a test item discriminates between a person who does well on the overall test versus a person who does poorly on the overall test. Item bias across samples is thought to exist if there is a significant difference between the measures of item difficulty or item discrimination as they are calculated from one sample of examinees and those same measures calculated from a second sample.

In order to illustrate the problems that arise with the classical test theory approach to item bias, let us consider the concept of item difficulty as it is applied to the following example. If the same test item of medium difficulty were administered to two groups of examinees, one of which was composed of high ability examinees and one of which was composed of low ability examinees, then the test item would be shown to be difficult in the case of the low ability group and easy in the case of the high ability group. Hence, the classical measure of item difficulty describes the difficulty of the item in relation to the group of examinees being tested; i.e., the classical measure of item difficulty is not independent of the frequency distribution of scores within the sample. A parallel argument may be presented with regard to item discrimination. Because item bias is confounded with sample differences, the traditional methods by which item difficulty and item discrimination are computed, may be inappropriate in order to assess item bias accurately.

An alternative to the classical test theory approach to the measurement of item bias is provided by latent trait theory (Lord, 1952, 1953; Lord & Novick, 1968). As latent trait theory works from an assumption of sample invariance, the item parameters (discrimination and difficulty) are independent of group performance. Thus, test bias or item bias may be assessed on the assumption that those subjects in different samples who have the same ability level have the same probability of getting a test item correct. This sample invariant property will then allow an investigator to assess test or item bias more accurately than the more traditional methods.

Purpose

The purpose of this thesis was to demonstrate an application of latent trait theory which assessed whether or not selected tests, that were employed in experimental designs requiring the evaluation of test-retest data or the testing of two or more different groups of examinees with the same test, were measuring the same latent trait(s) each time that they were employed. This, in turn, allowed the assessment of the validity of making a comparison between the different test results. The basis on which the comparison of the data between groups was made was derived from the premise that those test items that were not appearing to be measuring the same underlying trait(s) in both test administrations were not to be included in the final comparisons.

In order to illustrate this application of latent trait theory, a portion of the data that were collected for the Edmonton Grade III Achievement Study (Clarke, Nyberg, & Worth, 1977) was employed in this thesis. The Edmonton Grade III Achievement Study incorporated a battery of tests that was administered to all the Grade III students in the Edmonton Public School system in 1956. An almost identical battery of tests was administered to all the Grade III students in the Edmonton Public School system in 1977. Thus we have a situation where the same tests were administered to two groups of examinees, 21 years apart. A co-purpose, then, was to determine, through latent trait theory, whether or not the results of the tests of these two different test administrations could be directly compared with each other in order to assess achievement differences, between the two groups when they were tested 21 years apart. Furthermore, since some of the test items employed in the Edmonton Grade III Achievement Study were found to be favoring one or the other group of students, then the achievement differences between the two study years were calculated when these data were removed prior to making the comparisons.

In addition, Clarke et al. (1977) reported the item statistics (item difficulty and item discrimination) of the items used in the Edmonton Grade III Achievement Study that were calculated by classical test theory methodology. Although the item statistics that were calculated by classical test methodology were not directly comparable to the item statistics derived from latent trait theory, these results did allow a comparison of the implications of the results of the two different approaches to be made.

In summary, then, this thesis:

1. Demonstrated an application of latent trait theory for determining item bias in tests that are administered on different occasions. This, in turn, allowed for the determination of the validity of comparing data obtained 21 years apart in order to assess achievement differences between the different groups of examinees.
2. Determined the ability differences between the 1956 and

the 1977 Edmonton Public School Board grade III students, as measured by the application of latent trait theory, on a portion of the data obtained from the Edmonton Grade III Achievement Study.

3. Provided a discussion of the differences of the implications of the test statistics that were computed from these data via classical test theory methodology and latent trait theory methodology.

The following chapter provides a review of related literature as it pertains to (a) latent trait theory and (b) the Edmonton Grade III Achievement Study.

II. REVIEW OF RELATED LITERATURE

A. Latent Trait Theory

A general theory of latent traits presupposes that an individual's behavior can be predicted or explained by defining human characteristics, called traits, and that a person's performance on any given trait is based on the ability s/he possesses on that underlying trait (Lord & Novick, 1968). As these traits are measured indirectly, they are referred to as being latent. Hence, the observed examinee's test result and the underlying trait or ability should be related.

In latent trait theory, the relationship between item performance and ability is expressed by the item characteristic curve; a plot of the probability of getting the item correct as a function of the ability level expressed in ability units.

An important property of latent trait theory is that estimates of item characteristic curves do not depend on the invariant distributions of a given ability across different groups of examinees (Birnbaum, 1968). In fact, it need not be necessary to use the identical test forms in order to make valid comparisons across different groups of examinees, as long as the tests measure the same traits. These properties should allow researchers to assess changes of construct validity across time for different groups of examinees. In deriving the working mathematical model for latent trait analysis, Birnbaum (1968) outlines three fundamental notions that apply in general to latent trait theory: The dimensionality of the latent space, local independence, and item characteristic curves.

Dimensionality

Any examinee's test score can be expressed as being represented by k latent traits. If the examinee's test score is determined by a single ability, then the trait is unidimensional. If the examinee's test score is represented by more than one ability, then the test is said to be multidimensional. These k traits can be said to form a k dimensional space in which the examinee can be positioned. Most tests attempt to measure a single ability or trait. Consequently, many latent trait models make the assumption of unidimensionality of a test. Factor analysis is usually advocated to assess the dimensionality of a given test (e.g., Hambleton & Traub, 1973; Lumsden, 1976) in order to ascertain whether a unidimensional or multidimensional latent trait model would be more appropriate for a given test. However, Rentz and Rentz (1979) suggest that

> factor analysis does not clarify the dimensionality because factor analysis is itself a model (or a number of models) with several, sometimes conflicting, concepts of dimensionality.

The problem is further compounded when a reseacher is interested in the performance of a test administrated on two separate occasions; i.e., if there is a difference in the composition of the samples employed, e.g., if one sample was heterogeneous and another sample was homogeneous, then the resulting dimensionalities as measured by factor analysis may be different due to sample configuration and not due to the way in which the items were measuring the underlying trait. Thus, there may be no clear definition of undimensionality beyond the mathematical definition (Rentz & Rentz, 1979). If the test were measuring the same underlying trait each time that it was administered, then it would be assumed that the test dimensionality would be the same.

Local Independence

Local independence means that the item scores are related to each other only through the latent variables that are being measured; i.e.,

> within any group of examinees all characterized by the same values θ , θ , :.. θ , the (conditional) distributions of the item scores are all independent of each other. (Lord & Novick, 1968)

This implies that for a fixed ability level, the joint probability of the performance of item i and item j equals Pi x Pj when these items are locally independent; e.g., if an examinee had a probability of .7 of answering item i correctly and a probability of .8 of answering item j correctly, then the examinee has a probability of (.7) x (.8) = .56 of passing both items. Similarly, if one examinee had a probability of .4 of answering item i correctly and another examinee of the same ability had a probability of .5 of answering item i correctly, then the probability of of both examinees answering item i correctly will be (.4) x (.5) = .20. This also implies that for a fixed ability level, the probability of an examinee answering item j correctly is not changed by his success on item i. If this were not the case, then more than one trait would be needed in order to explain differences that might arise across different examinees of equal ability. In essence this can only be achieved if all of the test items measure a single trait. Hence, the techniques used to measure unidimensionality are also appropriate to establish whether or not the test items are locally independent, but they are also subject to the same problems.

Item Characteristic Curve

As noted earlier, the mathematical function that relates the probability of achievement on a test item to the ability measured by the test on the whole, is known as the item characteristic curve. This relationship takes the form of a nonlinear regression function of the item score on the latent trait represented by the test (Hambleton & Cook, 1977). An important property of the item characteristic curve is that the distribution of ability within a given group of examinees will not affect the shape of the item characteristic curve. Thus, if the item characteristic curve for two groups of examinees differed when being assessed on the same underlying trait, then the differences could not be explained in terms of differences in ability; but rather that the item is not measuring the same trait in the same way in both cases.

Hence, the item characteristic curve is the prime instrument that can be used to establish whether or not an item is performing in the same manner each time that it is administered. When a researcher compares the means of each sample that s/he obtained on a given test administration, then s/he is assuming that each test item is performing in the same manner on each of those test administrations and that the mean difference between those samples reflects a true difference of achievement. While this assumption is rarely tested, it would appear that, as the sample varies in composition and/or as the inter-test interval increases, then the probability increases that the test items may not be performing in the same manner on both test administrations. Consequently, the mean achievement differences between samples may not reflect a true achievement difference, but rather an achievement difference confounded by differences in the performance characteristics of the instrument; i.e., item bias. It would appear that the best way to assess item bias that arises when a test is administered on two or more occasions lies within latent trait theory.

Different Types of Latent Trait Models

Latent trait models differ from each other in the number of item parameters used to produce the item characteristic curve and with respect to the assumed fundamental relationship between ability and the probability of success. Lord (1952, 1953) developed a latent trait model where the item characteristic curve takes the form of the normal ogive. An examinee's performance was related to ability and to two item parameters: item difficulty and item discrimination. Birnbaum (1968) altered this model by using the logistic cumulative function instead of the normal-ogive function. Birnbaum also added a third item parameter (c) in order to account for the probability of guessing a correct response on a given item for an examinee of a given ability level.

The one-parameter logistic model, known as the Rasch model (Rasch, 1966) is recognized as a special case of Birnbaum's two-parameter logistic model. The Rasch item characteristic curves are calculated under the condition where all items are considered to have equal discriminating power and where guessing is not considered to be a factor; The sole parameter is a difficulty parameter[2] (Wright, 1968). The Rasch model has some advantages over the two- or three-parameter models because it is easier to compute the examinee's abilities and the computational problem of trying to estimate a larger number of different parameters is avoided. However, the degree of robustness of the Rasch model may play an important role in deciding the appropriateness of its application (Hambleton & Traub, 1971).

Samejima (1969, 1972, 1973) has introduced several different types of latent trait models to accommodate those tests that are not scored in the dichotomous fashion required by the different latent trait models that have been discussed so far. The nominal response model introduced by Bock (1972) and Samejima (1972) can be used when a given test is scored multichotomously as opposed to dichotomously; i.e., there are a number of alternatives from which the examinees may make their selection. The nominal response model allows for the calculation of "item option characteristic curves" for each respective item option. If the item responses of a given test can be ordered; i.e., the alternatives can be ordered in terms of correctness, then the graded response model (Samejima, 1969) would be able to accommodate the different type of item responses by producing an "operating characteristic" based upon the two-parameter logistic model. A variation of the graded response model is the continuous response model (Samejima, 1973). This model was introduced by Samejima in order to accommodate those tests that require the flexibility of allowing the examinee to respond on a continuum. This type of response often occurs with studies investigating affective traits.

---

[2] The Rasch model was developed independently of Birnbaum's model and Rasch is usually given credit for recognizing the usefulness of a one-parameter model.

While these different latent trait models were developed to accommodate a particular data configuration, each model is based on the general form:

$$P_i(\theta) = f(b_i, a_i) \qquad (1)$$

where:

$P_i(\theta)$ = the probability of getting item i correct for ability level θ;

b_i = the discrimination parameter for item i;

a_i = the difficulty parameter for item i.

Thus, the probability of a correct response by an examinee with a given ability on item i, is described by the item parameters a (discrimination) and b (difficulty). The different models may incorporate a different number of parameters, but each model will define precisely what the probability of success will be on a given test item for a particular ability level if the item parameters are known. For the three-parameter logistic model, the function is

$$P_i(\theta) = c_i + (1 - c_i)\left[\frac{1}{1+e^{-Da_i(\theta-b_i)}}\right] \qquad (2)$$

where:

c_i = the guessing parameter for item i;

D = 1.7.

For the two-parameter logistic model, the lower asymptote parameter (c) has a value of zero.[3] For the one-parameter logistic model (also known as the Rasch model) the discrimination (a) parameter is a constant. The problem is that these parameters are not known, so they must be estimated.

Methods of Obtaining the Item Characteristic Curve Parameters

There does not appear to be any one method of obtaining the item characteristic curve parameters that can be used with all testing situations. Consequently, several methods have evolved by which the item parameters may be estimated. In selecting an estimation procedure, the researcher must decide on the number of item parameters which s/he wishes to work with and the method by which the test items are to be scored. If the test items discriminate equally and guessing does not appear to be a factor, then the Rasch model would be the most appropriate model to use. If the test items were scored dichotomously and did not discriminate equally, then the two-parameter logistic model would be appropriate. If guessing was also a factor, then the three-parameter logistic model would be an appropriate model to choose. If there were several alternative answers that the examinee could choose from, then the researcher could consider one of the nominal response models; i.e., the graded response model if the alternatives could be ordered for correctness, or the continuous response model if the items are marked on a continuum. Having selected the model that the researcher is going to follow, then the examinee abilities and the item parameters must be estimated.

---

[3] This third parameter (c) is sometimes referred to as the guessing parameter.

The maximum likelihood method of estimating the item parameters has been the most common approach used to date. This method usually assumes a unidimensional latent space that has been defined by test items that have been scored dichotomously. Birnbaum (1968) points out that (i x n) + (N - 2) likelihood equations need to be solved; (where i = the number of item parameters, n = the number of test items, and N = the number of examinees); in order to obtain the maximum likelihood estimates of the item parameters. The method suggested by Birnbaum (1968) to solve the equation set involves estimating initial values for the item parameters and the ability estimates in the likelihood equation set and adjusting them in small steps to fit the data. The estimation procedure is considered converged when the difference between the successive item parameter estimates becomes less than the errors of calculation, causing the criterion function to fluctuate (Wood, Wingersky, & Lord, 1976). Different methods by which the initial item parameter estimates are made, the type of algorithm used to produce the item parameters, and the type of constraints that are placed on that algorithm in order to provide the best solution have been advocated by different researchers (e.g., Birnbaum, 1968; Bock, 1972; Anderson, 1970; Wright &

Douglas, 1977). The issues that are being debated by these and other researchers arise from the desired objective of finding an efficient algorithm in relation to the final item parameters that are estimated. The iterative process is very time consuming to take to completion and is, therefore, costly. A summary of the different methods used to estimate the item parameters was presented by Hambleton et al. (1977). The iterative procedure needed to obtain a solution necessitates the use of a computer.

The LOGIST computer program written by Wood, Wingersky, and Lord (1976), computes estimates of examinee ability and item charateristic curve parameters using an iterative process described as a modified Newton's method. The test item parameters are estimated via the maximum likelihood methods described by Lord (1968, 1974) which, in turn, are based upon Birnbaum's three-parameter logistic model (Birnbaum, 1968). If a researcher were interested in estimating the item parameters according to the Rasch model using a corrected unconditional maximum likelihood procedure, then BICAL, written by Wright and Mead (1976), would be an appropriate choice. One of the problems that a researcher experiences at this time is the lack of a wide distribution and/or selection of computer programs dealing with latent trait theory. However, the amount of information on latent trait theory is steadily growing (e.g., the summer of 1977 issue of the Journal of Educational Measurement dealt entirely with the subject of latent trait theory) and the influence of latent trait theory is spreading within the research community.

Applications of Latent Trait Theory

The majority of the literature reported on the subject of latent trait theory

> has been intended for measurement theorists, without hints on practical application, highly symbolic and, perhaps concise to the point of obfuscation. (Jaeger, 1977)

However, the practical application of latent trait theory is emerging; primarily due to the availability of new computer programs (e.g., Wood, Wingersky, & Lord; 1976) and the high speed computers necessary to run them. Hence, the bridge between theory and application is being built and is forming the basis of "a revolution . . . in the field of educational measurement" (Marco, 1977). This revolution is occurring in response to some of the problems that have been associated with classical test theory (see Lumsden, 1976).

Thus far, latent trait theory has been demonstrated as being able to provide acceptable methodology in approaching many stastistical problems (see Lord, 1977). In summary, Lord (1977) reports that item characteristic theory

> provides us with the frequency distribution f of test score(s) for examinees having a specific level of ability or skill.

This, then expands the information that is normally obtained from classical test theory methodology.

Lord (1977) further points out that

> if we have a pool of pretested items, all measuring the same trait or ability, we can predict the mean, variance, reliability, and raw score frequency distribution of any test constructed from these items once we know the ability levels in the group to be tested.

Individualized or tailored tests are made possible, then, by estimating the examinee's ability and then providing a set of questions that will maximize the information that can be obtained.[4] Item banking then becomes viable when latent trait theory is applied to the selection of items that constitute any individual test (Urry, 1977).

Latent trait theory also provides for the differential weighting of response alternatives. The nominal response model (Samejima, 1972; Bock, 1972) has been applied to those tests where the options to each item can be weighted with regard to degree of correctness (e.g., Thissen, 1976).

Another application of latent trait theory emerged when Lord (1973) estimated the power scores for a group of examinees who had taken a mistimed verbal aptitude test. This application of latent trait theory was made possible because the examinees had answered sufficient items to allow the estimation of the examinee's ability. Once this had been accomplished, the probability of the examinee getting the remaining items correct was calculated. An estimate of the examinee's power score was then obtained by summing the probable scores on the unanswered items and the examinee's obtained score. This procedure was possible only because latent trait theory provides the methodology to determine the probability of an examinee getting a correct score for a given ability level.

---

[4] This is especially true of the information obtained at the extremes of the ability distribution (Hambleton et al., 1977)

In addition, latent trait theory has been applied to test data in order to determine the equivalence of different forms of tests measuring the same trait. Lord (1975) has reported a summary of the different methods by which this can be achieved. However, the basic procedure involves the equation of the student ability estimates across the different tests that are pruported to be measuring the same ability.

Latent trait theory has also been applied to the measurement of item bias. While the concept of test bias has been an underlying concern of past researchers, the general social concern about such issues is also rising in connection with the testing and the classification of minority groups or the competency evaluation of students (e,g., Angoff, 1975).

In latent trait theory, if the test items are measuring the same latent trait each time a given test is administered then the item parameters will be linearly related across the different groups of examinees. This approach has been used with both the three-parameter logistic model (e.g., Lord, 1976a) and the Rasch model (e.g., Wright, Mead, & Draba, 1976).

Wright, Mead, and Draba (1976) report having applied the Rasch model to compute goodness of fit residuals for between-group differences in order to assess item bias. In order for this approach to be made feasible, however, the data should meet the requirements of the Rasch model. As mentioned earlier, however, some researchers report that the robustness of this model has not yet been fully tested (e.g., Hambleton & Traub, 1971).

Lord (1976a) reported the beginning stages of the development of an asymptotic significance test based on the summed variance - covariance matrices of the difficulty (b) and the discrimination (a) parameters. But, as was reported by Lord (1976a),

> it is not presently possible to specify with certainty . . . the asymptotic standard error of the maximum likelihood estimates used.

A third approach to measuring item bias has been reported by Rudner (1977). In this study, he computed the area of the differences between the equated item characteristic curves. This method appears to be the most feasible approach of measuring item bias by latent trait methodology, to date, but the range of abilities (-5.00 to +5.00) from which he calculates the area of difference between the two item characteristic curves may be extreme: i.e., he does not account for the difference between the number of examinees found at the extreme ability levels (which will be low) and the number of examinees found at the medium ability levels (which will be high).

While researchers are beginning to apply latent trait analysis procedures to some research projects, there are many areas that are open to investigation. For example, nowhere in the literature has any researcher reported the investigation of item bias, using latent trait methodology, of a test that was administrated to two groups of examinees, 21 years apart, in order to assess achievement differences between the two groups.

B. Edmonton Grade III Achievement Data

The Edmonton Grade III Achievement Study was conducted for the Alberta Advisory Committee on Educational Studies by S.C.T. Clarke, V. Nyberg, and W.H. Worth.[5] In conjunction with a team of University of Alberta students, the author was responsible for the data collection, test marking, data coding, and data analysis for the Edmonton Grade III Achievement Study.

Due to the complexity and relatively large size of this project (involving over 8,000 subjects and 285 test items), the data collection and processing were accomplished only through the co-operation of a large number of students, teachers, principals, and their respective schools; as well as the Edmonton Public School Board, the Department of Education for the province of Alberta, the Faculty of Education of the Univerity of Alberta, and the original investigators of the 1956 study.[6]

---

[5] N.M. Purvis was also involved with the Edmonton Grade III Achievement Study until his death in March, 1977.

[6] Many people were associated with the 1956 Edmonton Study of Achievement (e.g, G.M. Dunlop, S.C.T. Clarke, R.S. Mac Arthur, R.C. Harper, & W.H. Worth). The results of

The man-years of work that were expended in the acquisition of these data are large, but undoubtedly provide a wealth of information on the differences of student achievement after a period of 21 years.

The following five tests made up a comprehensive test battery that was administered to the Grade III students of the Edmonton Public School system in 1956 and again in 1977:

1. Raven Progressive Matrices (1947). Sets A, Ab, B;
2. California Short Form Test of Mental Maturity (Sullivan, Clark, & Tiegs, 1953);
3. Gates Advanced Primary Reading Tests for Grade 2 (second half) and Grade 3, Form I (Gates, 1942);
4. Gates Advanced Primary Reading Tests, Paragraph Reading, for Grade 2 (second half) and Grade 3, Form I (Gates, 1942);
5. California Achievement Tests Complete Battery, Primary Grades 1, 2, 3, and 4, Form CC, Reading, Arithmetic, and Language (Tiegs & Clark, 1950).

The Raven Progressive Matrices Test, Sets A, Ab, and B, contain a total of 36 items. Each item presents a figure with a section of the figure missing. The examinee must select the correct missing section from the six alternatives presented for each item. The test is untimed and is described by Raven (1965) as "a test of observational and clear thinking". The 1962 edition of this test is still in

---

[6] (cont'd)this study were reported in various minutes of the Alberta Advisory Committee on Educational Research and the minutes of the Edmonton Public School Board (Clarke et al., 1977).

print as of 1979, and may be obtained commercially.

This is the only test administered to the 1977 students that did not coincide exactly with that administered in 1956. The 1956 students were administered the 1947 edition of this test and the 1977 students were administered the 1962 edition. The differences between the two editions are slight and as the 1947 edition was not available in 1977, the decision was made to use the 1962 edition. The test items in both editions (1947; 1962) are identical, but in subtest A, the order of items 11 and 12 was interchanged. In addition, the order of some of the distractors was rearranged for a few of the items in the 1962 edition. The placement or order of the correct distractor was, however, not changed for any item. As this test was scored dichotomously, the exact order of the distractors did not play a part in the LOGIST analysis.

The California Short Form Test of Mental Maturity is out-of-print, but the publishers gave permission to the investigators of the Edmonton Grade III Achievement Study to reprint the test. A copy of this exam is presented in Appendix A. The California Short Form Test of Mental Maturity assesses the general intellectual maturity of students using 98 questions that vary in composition and complexity. While time limits are imposed on the student, the test is designed as a power test and not as a speed test. The examiners administered the directions and questions for each test item as per the instructions in the test manual. A copy of the directions for administration of this test is presented in Appendix A.

The Gates Advanced Reading Test: Word Recognition is out-of-print. The publishers granted permission for the investigators of the Edmonton Grade III Achievement Study to reproduce the test. A copy of this test is presented in Appendix A. The Gates Word Recognition Test was designed to assess a student's ability to select a word that best describes an accompanying picture from among four alternatives for each of 48 items. The test items vary in difficulty, from easy to hard, and were constructed under the premise that, "The fewer [items] he [the student] can recognize without error, the less ready he is to do independent reading" (Gates, 1943).

The Gates Advanced Primary Reading Test: Paragraph Reading is out-of-print. The publishers granted permission to the investigators of the Edmonton Grade III Achievement Study to reproduce this test and a copy of the test is presented in Appendix A. The Gates Paragraph Reading Test was designed to assess a student's ability to read and understand a statement directing the students to perform a particular task and to carry out that specified directive. Or, as described by Gates (1943), "This test measures ability to read thought units with full and exact understanding of the whole." This test is composed of 16 one-directive questions and 8 two-directive questions. The tests are timed, but in general, the time allotted allows all students to attempt all questions. Thus, this test was not designed to be a speed test.

The California Achievement Tests Complete Battery (Tiegs & Clark, 1950) is a comprehensive achievement test containing several sub-tests covering a wide range of abilities. While this test was administered to the 1956 students and the 1977 students, these data were not included in the data analyzed in this thesis since the appropriateness of the application of latent trait theory could be adequately investigated by using the data obtained from the other four tests.

Data Collection Procedures

The testing took place about the end of May, both in 1956 and 1977. The methods by which the data were collected in 1956 were again followed in 1977. The achievement tests were administered in the classroom by the teachers, and the mental maturity tests were administered by a team of University of Alberta students. The data from 1956 had been retained and were re-marked and coded for computer analysis at the same time that the 1977 data were processed. A team of University of Alberta students was responsible for the test marking and data coding. The data were then transferred to magnetic tape by the department of education, Alberta Education. A full description of the data collection was reported by Clarke, Nyberg, and Worth (1977, 1978).

The purpose of the 1956 study was to assess the degree of achievement of the Edmonton Public School Grade III students; while the etiology for the 1977 study was to assess the difference in achievement between the 1956 Edmonton Public School Grade III students and the 1977 Edmonton Public School Grade III students. There were approximately 3,500 students involved in the 1956 study and 4,500 students involved in the 1977 study. To cause the least inconvenience to the co-operating teacher, the battery of tests was administered over a number of days. No attempt was made to recover data that were lost as a result of student absenteeism. Consequently, the number of students, that wrote any one test, varied.

C. Chapter Summary

The advantages of latent trait theory methodology over classical test theory methodology has been advocated by many researchers (e.g., Lord & Novick, 1968), although the latent trait methodology has not yet been widely applied. The importance of latent trait methodology in determining item bias has also been demonstrated (e.g., Rudner, 1977), but a method of determining item bias through latent trait methodology which also accounts for the frequency differences among the examinee ability levels has not been demonstrated. In addition, the application of latent trait theory has not been applied to test data in order to determine the validity of making achievement score comparisons of tests that were administered many years apart. This thesis, then, attempts to demonstrate an application of latent trait theory in order to determine item bias without having to ignore the frequency differences between the different examinee ability levels.

In demonstrating this application of latent trait theory, the validity of comparing a portion of the 1956 data with the 1977 data from Edmonton Grade III Achievement Study was determined along with the resulting achievement differences between the two groups. In addition, it was then possible to discuss the implications of the latent trait theory item statistics with the implications of the classical test theory item statistics reported by Clarke, Nyberg and Worth (1977). The following chapter describes the method by which the data were analyzed.

III. METHOD

Data Analysis

The data from the Edmonton Grade III Achievement Study were recoded to the input format required by LOGIST (Wood, Wingersky, & Lord, 1976). The University of Alberta's Division of Educational Research Services' scoring program (SCOR01) was employed to score the data, which were then subsequently reformatted as documented by Wood et al. (1976, p. 1). Each test from each study year (1956 or 1977) was then analyzed by LOGIST and the examinee abilities (THETAs) and the item characteristic parameters were estimated.

The item parameters on a particular test administered in 1956 can not be directly compared to the item parameters on the same test administered in 1977. As Lord (1976a) indicates, this

> is inherent in the nature of the problem that the origin and the unit for measuring ability cannot be determined from the data.

Before a comparison of the item characteristic curves can be made between the two study years the item characteristic curve parameters that were obtained from the 1977 students must be rescaled to the scale obtained from the 1956 students. The only term in the item characteristic function (2) that is dependent on the examinee ability value (θ), the difficulty (b), and discrimination (a) item parameters is a (θ - b). The scale values of these variables can be transformed linearly, but the transformation must be accomplished without changing the prediction of the latent trait model which, in this case, is represented by function (2) (Allen & Yen, 1979; pp 256): i.e., it is necessary that

$$a_i^*(\theta^* - b_i^*) = a_i(\theta - b_i) \qquad (3)$$

where:

* denotes that the value has been rescaled.

In order to comply with the constraint set forward by formula 3, the 1977 item characteristic curve parameters were rescaled to the 1956 item parameters in the following manner. The 1977 item parameters were rescaled by using items that the logistic model fit well. An item was considered to be ill fitting, or poorly defined if it had a difficulty (b) parameter that was greater than 4.00 or less than -4.00 on either the 1956 data or the 1977 data. Thus, only those items that had a difficulty (b) parameter that was within the range of plus or minus 4.00 were used for the purpose of rescaling the 1977 data.

This procedure was necessary in order to prevent those items that were extremely easy or extremely hard from over-influencing the rescaling procedure. If the items that had an extreme difficulty (b) parameter were not eliminated, then the rescaled parameter would not fit the logistic model and equation 3 would not be balanced.

Another reason necessitating this procedure was that the item parameters were obtained after only 16 iterations of the LOGIST program and not necessarily at convergence. This decision was based on a cost-benefit estimate of allowing LOGIST further computer time. As the data from both the 1956 and 1977 students were analyzed in the same manner, and as those items with extreme difficulty (b) parameters had little to offer in the way of test information; there was little reduction of test information and a large saving in computing expenses.

Those items that had a difficulty (b) parameter ≤ -4.00 or ≥ 4.00 made up the set of items that formed the basis of the standardization procedure. The mean and standard deviation of this set of items were calculated for the difficulty (b) parameter from the 1956 data. The item difficulty (b) parameters from the 1977 data were then rescaled to the scale obtained from the 1956 data by the following formula:

$$b^*_{i_{77}} = \overline{X}_{b_{56}} + (Z_{b_{i_{77}}} \cdot s_{b_{56}}) \qquad (4)$$

where:

$b^*_{i_{77}}$ = the rescaled 1977 difficulty parameter for item i;

$\overline{X}_{b_{56}}$ = the mean of the 1956 difficulty parameter;

$Z_{b_{i_{77}}}$ = the 1977 difficulty parameter Z score for item i;

$s_{b_{56}}$ = the standard deviation of the 1956 difficulty parameter.

The discrimination (a) parameter must be rescaled in order to maintain the same amount of information that was obtained by the original discrimination (a) parameter in relation to the original difficulty (b) parameter. This may be visualized by noting that the discrimination (a) parameter controls the slope of the item characteristic curve and, if the difficulty (b) parameter is rescaled, then the slope of the curve is going to have to be adjusted accordingly if equation 3 is going to remain true.

The 1977 data item discrimination (a) parameters were rescaled by the following formula:

$$a^*_{i_{77}} = \frac{s_{b_{77}}}{s_{b_{56}}} \cdot a_{i_{77}} \qquad (5)$$

where:

$a^*_{i_{77}}$ = the rescaled 1977 discrimination parameter for item i;

$s_{b_{77}}$ = the standard deviation of the 1977 difficulty parameter;

$a_{i_{77}}$ = the original 1977 discrimination parameter for item i.

If the 1977 data item parameters are rescaled then the 1977 student abilities must also be rescaled. Given that the item difficulty (b) and the ability scale are the same scale, then the student abilities must be rescaled according to the following formula:

$$\theta^*_{77} = \left[\frac{\theta_{77} - \overline{X}_{b_{77}}}{s_{b_{77}}} \cdot s_{b_{56}} \right] + \overline{X}_{b_{56}} \qquad (6)$$

where:

θ^*_{77} = the rescaled 1977 ability level;

θ_{77} = the original 1977 ability level.

The rescaled value of ability will allow the determination of the distribution of the rescaled 1977 student abilities in relation to the 1956 student abilities.

The lower asymptotic (c) item parameter of the 1977 data was reset to the respective 1956 data levels. This was necessary because the logistic model indicates that if the difficulty (b) parameter and the discrimination (a) parameter are rescaled then they must adopt the lower asymptote of the new scale (Marco, 1977).

The following proof is offered as substantiation that equation 3 is true when the rescaling formulae 4, 5, and 6 are applied to the rescaling of the 1977 item parameters to the scale obtained from the 1956 data.

$$a^*_{i_{77}}(\theta^*_{77} - b^*_{i_{77}}) = \frac{s_{b_{77}}}{s_{b_{56}}}(a_{i_{77}}) \cdot \left\langle \left[\left(s_{b_{56}} \cdot \frac{\theta_{77} - \bar{X}_{b_{77}}}{s_{b_{77}}} \right) + \bar{X}_{b_{56}} \right] - \left[\left(s_{b_{56}} \cdot \frac{b_{i_{77}} - \bar{X}_{b_{77}}}{s_{b_{77}}} \right) + \bar{X}_{b_{56}} \right] \right\rangle$$

$$= \frac{s_{b_{77}}}{s_{b_{56}}}(a_{i_{77}}) \cdot \frac{s_{b_{56}}}{s_{b_{77}}} \cdot \left[(\theta_{77} - \bar{X}_{b_{77}}) - (b_{i_{77}} - \bar{X}_{b_{77}}) \right]$$

$$= a_{i_{77}}(\theta_{77} - b_{i_{77}}).$$

Using the rescaled parameters, the probabilities of success for each ability in the range of plus or minus 3.00 in steps of 0.2 were calculated from equation 2. The item characteristic curves were then plotted.

A probable difference of item scores was calculated by multiplying the differences (between 1956 & 1977) of the probability of success for a given ability level by the proportion of total scores found at that ability level, then summing across all ability levels, i.e.,

$$P_i(\delta) = \sum_{\theta=-3}^{\theta=+3} \left[P_{56}(\theta) - P_{77}(\theta) \right] \cdot \frac{n_{56} + n_{77}}{N_{56} + N_{77}} \qquad (7)$$

where:

$P_i(\delta)$ = the probable difference of item score between the 1956 and 1977 students on item i;

$P_{56}(\theta)$ = the probability of success in 1956 for a student at a particular ability level θ;

$P_{77}(\theta)$ = the probability of success in 1977 for a student at a particular ability level θ;

n_{56} = the number of 1956 examinees having the ability level θ;

n_{77} = the number of 1977 examinees having the ability level θ;

N_{56} = the total number of examinees for 1956;

N_{77} = the total number of examinees for 1977.

Those items that had a probable difference of item scores greater than or equal to 0.05, 0.10, or 0.15 were identified as being aberrant.

In order to determine the differences in abilities between the 1956 group of examinees and the 1977 group of examinees, the data from both study years were pooled and subsequently analyzed by LOGIST. Because the resulting abilities were then on the same scale, they were directly comparable. Consequently, the abilities from the 1956 students and the 1977 students were identified and the mean and standard deviation of the respective abilities were then computed for each study year for each test. The difference between these means represents the mean ability difference between 1956 and 1977 on any given test.

In order to determine the effects of the aberrant items at each difference level, the above procedure was repeated after the appropriate aberrant items were removed. Thus, mean student ability differences between the 1956 students and the 1977 students were obtained under the following conditions:

1. when no test items were removed,
2. when those test items that were aberrant at the ≥.15 difference level were removed,
3. when those test items that were aberrant at the ≥.10 difference level were removed,
4. when those test items that were aberrant at the ≥.05

difference level were removed.

As a technical aside, it should be noted that LOGIST deletes all those students having a 100% score from the sample before making its calculation of abilities. Consequently, all 100% scores were re-introduced as an ability of +4.00 before the mean differences of ability were calculated.

IV. RESULTS

The results chapter will be presented in five sections; one section for each test. The discussion of the results for each test takes place in the following chapter. Within each section the following information is presented:

1. The sample size of the 1956 Grade III study and the sample size of the 1977 Grade III study for the selected test.
2. When the test numbering sequence differs from the numbering sequence used in the analysis procedure, a table showing the relationship between the numbering of the original test and that used for the analysis procedure is presented.
3. The item parameters for the selected test as calculated by 16 iterations of LOGIST are presented for the 1956 Grade III data and the 1977 Grade III data.
4. The number of students that had obtained a particular ability level in 1956 and in 1977 are then presented. It should be noted that the sum of all the students across the ability levels -3.00 to +3.00 is less than the total number of students included in the analysis. This is because only the abilities between -3.00 and +3.00 were selected for the purpose of plotting the item characteristic curves. This was done in order to eliminate those instances where the results were poorly

determined due to the lack of subjects at the extreme values of ability. The difference then, between the total N calculated in this manner and the total N reported at the beginning of each test section is due to LOGIST's calculation of abilities outside the range of -3.00 to +3.00.

5. The rescaled item parameters for the 1956 data and the 1977 data for the selected test. This allows for the calculation of the item characteristic curves to be on the same scale, and, thus, be directly comparable.
6. The probable difference of scores between the 1956 students and the 1977 students for each item on the selected test are presented.
7. The average abilities of the 1956 Grade III students and the 1977 Grade III students are presented under the following conditions:
    a. for a set of test items that was constructed by removing all of the aberrant test items that had an item probable difference ≥ 0.05.;
    b. for a set of test items that was constructed by removing all of the aberrant test items that had an item probable difference ≥ 0.10.;
    c. for a set of test items that was constructed by removing all of the aberrant test items that had an item probable difference ≥ 0.15.;
    d. for the complete test, i.e., where no items have been eliminated prior to analysis by LOGIST.

Within each condition, the students that obtained a score of 100% are removed by LOGIST before computing the overall ability scores. Thus, the students who obtained a score of 100% were assigned an ability level of 4.00 before the ability means were calculated.

8. The item characteristic curves for the 1956 data and the 1977 data are presented together for each selected test item that was included in the standardization procedure prior to the rescaling of the item parameters obtained from the 1977 data. This was done in order to facilitate the comparison of the performance of the 1956 students and the 1977 students on a given item across all levels of student ability.

Raven Progressive Matrices Test

The data from 3,596 students who were administered Raven's Progressive Matrices Test in 1956 were analyzed along with the data from 2,577 1977 students. Table 1 presents the study item numbers used in the analysis and the related test item numbers. The 1977 sample represents approximately 60% of the Grade III students enrolled in the Edmonton Public School system in 1977. The remaining 40% of the 1977 population was not administered the Raven's Progressive Matrices Test. These students were randomly selected to receive an alternative test for a companion study that was being administered at the same time as the Grade III Achievement study.

The item parameters from the 1956 student data and the 1977 student data, when analyzed separately by LOGIST, are presented in Table 2. Table 3 presents the rescaled item parameters. The number of students found at each of the specified levels of ability are presented in Table 4.

The probable difference of scores between the 1956 Grade III students and the 1977 Grade III students for each test item are presented in Table 5. The results from the aberrant test items that were identified in Table 5 for each of the different aberrant levels were removed from the total test data. The selected data were then pooled across the two study years and were reanalyzed by LOGIST. This was done in order to obtain new student abilities which did not include the results from the aberrant items. Mean abilities were then calculated for both the 1956 Grade III students and the 1977 Grade III students at each of the specified levels of aberrance and are presented in Table 6. The item characteristic curves for the Raven Progressive Matrices Test are presented in Figures 2 to 26.

Gates Paragraph Reading

The data from 3,569 students who were administered Gates' paragraph reading Test in 1956 were analyzed along with the data from 4,430 1977 Grade III students. The test item number of the Gates Paragraph Reading test and the corresponding Grade III Achievement Study numbers are presented in Table 7.

The item parameters for the 1956 data and the 1977 data

Table 1

Study Item Number Assignment for Raven's Progressive Matrices Test

Raven(1947)	1956	1977	Raven(1947)	1956	1977
Item #	Study #	Study #	Item #	Study #	Study #
A 1	1	1	Ab 7	19	19
2	2	2	8	20	20
3	3	3	9	21	21
4	4	4	10	22	22
5	5	5	11	23	23
6	6	6	12	24	24
7	7	7			
8	8	8	B 1	25	25
9	9	9	2	26	26
10	10	10	3	27	27
11	11	12	4	28	28
12	12	11	5	29	29
			6	30	30
Ab 1	13	13	7	31	31
2	14	14	8	32	32
3	15	15	9	33	33
4	16	16	10	34	34
5	17	17	11	35	35
6	18	18	12	36	36

Table 2

Original Item Parameters for Raven's Progressive Matrices

	1956			1977		
ITEM #	a	b	c	a	b	c
1	0.129	-32.228	0.045	0.758	-5.140	0.105
2	0.266	-15.277	0.045	0.679	-5.477	0.105
3	0.395	-9.781	0.045	0.643	-4.532	0.105
4	0.260	-11.763	0.045	0.706	-4.079	0.105
5	0.327	-7.989	0.045	0.648	-3.934	0.105
6	0.293	-8.441	0.045	0.621	-3.991	0.105
7	0.493	-1.537	0.045	0.597	-1.879	0.105
8	0.289	-2.959	0.045	0.318	-2.122	0.105
9	0.502	-2.358	0.045	0.631	-1.946	0.105
10	0.476	-2.004	0.045	0.605	-1.617	0.105
11	0.537	1.550	0.045	0.683	0.617	0.105
12	0.492	0.432	0.045	0.672	1.358	0.105
13	0.561	-5.904	0.045	1.118	-3.283	0.105
14	0.441	-6.110	0.045	0.831	-3.159	0.105
15	0.240	-9.910	0.045	0.726	-3.204	0.105
16	0.738	-0.980	0.045	0.699	-1.728	0.105
17	0.655	-0.988	0.045	0.736	-1.237	0.105
18	0.996	-0.963	0.045	0.952	-0.961	0.105
19	0.790	-1.049	0.045	0.743	-1.485	0.105
20	0.767	-0.655	0.045	0.796	-0.559	0.105
21	0.817	0.268	0.045	0.824	-0.381	0.105
22	0.577	-0.180	0.045	0.612	-0.619	0.105
23	0.438	0.301	0.045	0.553	-0.339	0.105
24	1.939	1.068	0.056	1.343	0.809	0.142
25	0.368	-7.646	0.045	0.712	-3.768	0.105
26	0.553	-4.335	0.045	0.749	-3.130	0.105
27	0.833	-2.133	0.045	0.806	-2.337	0.105
28	0.931	-1.444	0.045	0.777	-2.058	0.105
29	0.908	-0.400	0.045	0.828	-0.814	0.105
30	1.035	0.268	0.068	0.720	-0.610	0.105
31	0.781	0.419	0.045	0.494	0.302	0.105
32	1.939	0.725	0.045	2.000	0.169	0.088
33	1.939	0.586	0.045	2.000	0.033	0.105
34	1.939	0.482	0.090	2.000	-0.256	0.173
35	1.939	0.709	0.029	2.000	0.016	0.065
36	1.939	1.133	0.016	1.179	0.978	0.043

Table 3

Rescaled Item Parameters for the Raven Progressive Matrices

	1956			1977		
ITEM #	a	b	c	a	b	c
1		N/A			N/A	
2		N/A			N/A	
3		N/A			N/A	
4		N/A			N/A	
5		N/A			N/A	
6		N/A			N/A	
7	0.493	-1.537	0.045	0.520	-1.781	0.045
8	0.289	-2.959	0.045	0.277	-2.059	0.045
9	0.502	-2.358	0.045	0.550	-1.857	0.045
10	0.476	-2.004	0.045	0.527	-1.480	0.045
11	0.537	1.550	0.045	0.585	1.936	0.045
12	0.492	0.432	0.045	0.595	1.086	0.045
13		N/A			N/A	
14		N/A			N/A	
15		N/A			N/A	
16	0.738	-0.980	0.045	0.609	-1.607	0.045
17	0.655	-0.988	0.045	0.641	-1.043	0.045
18	0.996	-0.963	0.045	0.829	-0.726	0.045
19	0.790	-1.049	0.045	0.647	-1.328	0.045
20	0.767	-0.655	0.045	0.693	-0.264	0.045
21	0.817	0.268	0.045	0.718	-0.061	0.045
22	0.577	-0.180	0.045	0.533	-0.333	0.045
23	0.438	0.301	0.045	0.482	-0.012	0.045
24	1.939	1.068	0.056	1.170	1.306	0.056
25		N/A			N/A	
26		N/A			N/A	
27	0.833	-2.133	0.045	0.702	-2.305	0.045
28	0.931	-1.444	0.045	0.676	-1.986	0.045
29	0.908	-0.400	0.045	0.721	-0.558	0.045
30	1.035	0.268	0.068	0.627	-0.324	0.068
31	0.781	0.419	0.045	0.431	0.723	0.045
32	1.939	0.725	0.045	1.742	0.571	0.045
33	1.939	0.586	0.045	1.742	0.415	0.045
34	1.939	0.482	0.090	1.742	0.083	0.090
35	1.939	0.709	0.029	1.742	0.395	0.029
36	1.939	1.133	0.016	1.027	1.500	0.016

N/A - This item was not included due to an extreme original difficulty parameter.

Table 4

The Number of Students Located at Each Level of Ability for the Raven Progressive Matrices Test

Ability	1956	1977*	Ability	1956	1977*
-3.00	23	3	0.20	367	92
-2.80	29	14	0.40	291	234
-2.60	34	25	0.60	245	253
-2.40	30	24	0.80	267	234
-2.20	46	12	1.00	240	107
-2.00	43	32	1.20	194	189
-1.80	66	46	1.40	97	134
-1.60	70	56	1.60	59	52
-1.40	89	30	1.80	23	96
-1.20	114	57	2.00	16	75
-1.00	133	63	2.20	8	49
-0.80	132	66	2.40	4	32
-0.60	177	49	2.60	11	5
-0.40	192	113	2.80	4	16
-0.20	225	149	3.00	1	15
0.00	311	177			

* The 1977 distribution following the rescaling procedure

Table 5

Aberrant Items at Specified Levels of Difference

Raven's Progressive Matrices Test

Item #	Diff.	≥ 0.05	≥ 0.10	≥ 0.15
1	N/A			
2	N/A			
3	N/A			
4	N/A			
5	N/A			
6	N/A			
7	0.0366			
8	0.0768	*		
9	0.0337			
10	0.0428			
11	0.0658	*		
12	0.1269	*	*	
13	N/A			
14	N/A			
15	N/A			
16	0.0645	*		
17	0.0062			
18	0.0621	*		
19	0.0258			
20	0.0862	*		
21	0.0714	*		
22	0.0228			
23	0.0512	*		
24	0.0514	*		
25	N/A			
26	N/A			
27	0.0122			
28	0.0348			
29	0.0315			
30	0.1186	*	*	
31	0.0798	*		
32	0.0529	*		
33	0.0586	*		
34	0.1255	*	*	
35	0.1079	*	*	
36	0.0647	*		

N/A - This item was not included due to an extreme original difficulty (b) parameter.

Table 6

Mean Student Abilities for Raven's Progressive Matrices Test

Level of aberrance		1956 students	1977 students	1956+1977 combined	# items removed	# items remaining
1.0	mean =	-0.204	0.276	-0.004	0	36
	s.d. =	1.060	1.189	1.140		
0.15		Same as above			0	36
0.10	mean =	-0.168	0.259	0.010	4	32
	s.d. =	1.662	1.237	1.158		
0.05	mean =	1.087	1.546	1.279	16	20
	s.d. =	2.039	2.172	2.108		
	N =	3596	2577	6173		

when analyzed separately by LOGIST are presented in Table 8. Table 9 presents the rescaled item parameters. The number of students that were located at each specific level of ability are given in Table 10.

The probable difference of scores between the 1956 Grade III students and the 1977 Grade III students for each test item are presented in Table 11. The results from the aberrant test items that were identified in Table 11 for each of the different aberrant levels were removed from the total test data. The selected test item data were then pooled across the two study years and reanalyzed by LOGIST. This was done in order to obtain new student abilities which did not include the results from the aberrant items. Mean abilities were then calculated for both the 1956 Grade III students and the 1977 Grade III students at each of the specified levels of aberrance and are presented in Table 12. The item characteristic curves for Gate's paragraph reading test are presented in Figures 27 to 49.

Gates Word Recognition

The data from 3,540 1956 Grade III students were analyzed along with the data from 4,430 1977 Grade III students. Table 13 presents the study item number and the related test item number. The item parameters from the 1956 data and those from 1977 data, when analyzed separately by LOGIST, are presented in Table 14. The rescaled item parameters are presented in Table 15. The number of students located at specified levels of ability are presented in

Table 7

Study Item Number Assignment for Gates Paragraph Reading

Item #	Study #	Item #	Study #
1	1	17a	17
2	2	17b	18
3	3	18a	19
4	4	18b	20
5	5	19a	21
6	6	19b	22
7	7	20a	23
8	8	20b	24
9	9	21a	25
10	10	21b	26
11	11	22a	27
12	12	22b	28
13	13	23a	29
14	14	23b	30
15	15	24a	31
16	16	24b	32

Table 8

Original Item Parameters for Gates Paragraph Reading

	1956			1977		
ITEM #	a	b	c	a	b	c
1	0.560	-5.193	0.075	0.499	-5.647	0.060
2	0.585	-5.446	0.075	0.238	-6.351	0.060
3	0.906	-3.989	0.075	0.562	-5.336	0.060
4	0.308	-7.354	0.075	0.430	-5.184	0.060
5	0.314	-6.098	0.075	0.301	-5.342	0.060
6	0.455	-5.070	0.075	0.380	-5.051	0.060
7	0.703	-1.698	0.075	0.525	-2.346	0.060
8	0.331	0.562	0.075	0.264	-1.198	0.060
9	0.638	-4.807	0.075	0.531	-5.936	0.060
10	0.581	-3.704	0.075	0.681	-3.313	0.060
11	0.322	-5.897	0.075	0.418	-4.617	0.060
12	0.563	-3.681	0.075	0.629	-3.499	0.060
13	0.465	-3.214	0.075	0.421	-3.076	0.060
14	0.767	-2.812	0.075	0.641	-2.946	0.060
15	0.612	-2.333	0.075	0.607	-2.274	0.060
16	0.468	-1.683	0.075	0.407	-2.312	0.060
17	0.747	-0.199	0.075	0.726	-0.299	0.060
18	0.362	-3.457	0.075	0.294	-4.725	0.060
19	1.003	-1.233	0.075	0.661	-1.196	0.060
20	0.839	-0.392	0.075	0.760	-0.277	0.060
21	2.000	-0.623	0.075	1.624	-0.952	0.126
22	1.598	-0.660	0.124	0.921	-1.251	0.060
23	1.224	-0.336	0.075	1.179	-0.153	0.001
24	0.655	-1.237	0.075	0.526	-1.159	0.060
25	0.758	1.060	0.233	0.789	0.666	0.199
26	0.778	1.432	0.155	0.879	0.974	0.133
27	2.000	0.633	0.040	2.000	0.473	0.037
28	1.092	0.926	0.027	0.957	0.745	0.042
29	0.715	1.101	0.236	0.728	0.998	0.283
30	2.000	0.959	0.063	2.000	0.971	0.060
31	2.000	1.522	0.075	2.000	1.173	0.060
32	0.745	1.436	0.027	0.818	1.059	0.042

Table 9

Rescaled Item Parameters for Gates Paragraph Reading

	1956			1977		
ITEM #	a	b	c	a	b	c
1		N/A			N/A	
2		N/A			N/A	
3		N/A			N/A	
4		N/A			N/A	
5		N/A			N/A	
6		N/A			N/A	
7	0.703	-1.698	0.075	0.488	-2.241	0.075
8	0.331	0.562	0.075	0.246	-1.007	0.075
9		N/A			N/A	
10	0.581	-3.704	0.075	0.634	-3.281	0.075
11		N/A			N/A	
12	0.563	-3.681	0.075	0.585	-3.481	0.075
13	0.465	-3.214	0.075	0.392	-3.026	0.075
14	0.767	-2.812	0.075	0.596	-2.886	0.075
15	0.612	-2.333	0.075	0.565	-2.164	0.075
16	0.468	-1.683	0.075	0.379	-2.204	0.075
17	0.747	-0.199	0.075	0.675	-0.041	0.075
18		N/A			N/A	
19	1.003	-1.233	0.075	0.615	-1.005	0.075
20	0.839	-0.392	0.075	0.707	-0.017	0.075
21	2.000	-0.623	0.075	1.511	-0.742	0.075
22	1.598	-0.660	0.124	0.857	-1.065	0.124
23	1.224	-0.336	0.075	1.097	0.117	0.075
24	0.655	-1.237	0.075	0.489	-0.965	0.075
25	0.758	1.060	0.233	0.734	1.996	0.233
26	0.778	1.432	0.155	0.817	1.328	0.155
27	2.000	0.633	0.040	1.860	0.790	0.040
28	1.092	0.926	0.027	0.890	1.081	0.027
29	0.715	1.101	0.236	0.677	1.354	0.236
30	2.000	0.959	0.063	1.860	1.325	0.063
31	2.000	1.522	0.075	1.860	1.541	0.075
32	0.745	1.436	0.027	0.761	1.419	0.027

N/A - This item was not included due to an extreme original difficulty parameter.

Table 10

The Number of Students Located at Each Specified Level of Ability for the Gates Paragraph Reading test

Ability	1956	1977*	Ability	1956	1977*
-3.00	17	6	0.20	272	321
-2.80	25	14	0.40	292	349
-2.60	28	20	0.60	255	350
-2.40	23	24	0.80	238	335
-2.20	31	33	1.00	207	325
-2.00	40	47	1.20	168	156
-1.80	53	77	1.40	112	265
-1.60	63	51	1.60	63	208
-1.40	77	120	1.80	38	116
-1.20	110	126	2.00	20	63
-1.00	150	154	2.20	24	50
-0.80	189	205	2.40	10	45
-0.60	209	215	2.60	7	22
-0.40	247	241	2.80	4	7
-0.20	249	128	3.00	1	13
0.00	285	286			

* The 1977 distribution following the rescaling procedure

Table 11

Aberrant Items at Specified Levels of Difference

Gates Paragraph Reading Test

Item #	Diff.	≥ 0.05	≥ 0.10	≥ 0.15
1	N/A			
2	N/A			
3	N/A			
4	N/A			
5	N/A			
6	N/A			
7	0.0293			
8	0.1540	*	*	*
9	N/A			
10	0.0068			
11	N/A			
12	0.0032			
13	0.0379			
14	0.0215			
15	0.0256			
16	0.0216			
17	0.0372			
18	N/A			
19	0.0961	*	*	
20	0.0884	*		
21	0.0294			
22	0.0685	*		
23	0.1200	*	*	
24	0.0778	*		
25	0.0127			
26	0.0114			
27	0.0467	*		
28	0.0304			
29	0.0315			
30	0.0854	*		
31	0.0045			
32	0.0022			

N/A - This item was not included due to an extreme original difficulty (b) parameter.

Table 12

Mean Student Abilities for the Gates Paragraph Reading Test

Level of aberrance		1956 students	1977 students	1956+1977 combined	# items removed	# items remaining
1.0	mean =	-0.137	0.017	-0.051	0	32
	s.d. =	1.159	1.164	1.164		
0.15	mean =	-0.133	0.006	-0.056	1	31
	s.d. =	1.670	1.179	1.175		
0.10	mean =	-0.180	0.002	-0.079	3	29
	s.d. =	1.195	1.214	1.209		
0.05	mean =	-0.132	0.057	-0.027	8	24
	s.d. =	1.158	1.217	1.195		
	N =	3569	4430	7999		

Table 16.

The probable difference of scores between the 1956 Grade III students and the 1977 Grade III students for each test item are presented in Table 17. The results from the aberrant test items identified in Table 17 for each of the different aberrant levels, were removed from the total test data. The selected test item data were then pooled across the two study years and reanalyzed by LOGIST. This was done in order to obtain new student abilities which did not include the results from the aberrant items. Mean abilities were then calculated for both the 1956 Grade III students and the 1977 Grade III students at each of the specified levels of aberrance and are presented in Table 19. The item characterisic curves for Gate's Word Recognition Test are presented in Figures 50 to 88.

California Mental Maturity

The data from 3,443 1956 Grade III students and the data from 4,378 1977 students were analyzed using LOGIST. Table 19 presents the study item numbers and the related test item numbers. The item parameters calculated from the 1956 data and those calculated from the 1977 data when they were separately analyzed using LOGIST are presented in Table 20. The rescaled item parameters are presented in Table 21. The number of students found at each specified level of ability are presented in Table 22.

The probable difference of scores between the 1956 Grade III students and the 1977 Grade III students for each

Table 13

Study Item Number Assignment for Gates Word Recognition

Item Word	Study #	Item Word	Study #
apple	1	stumble	25
floor	2	chirp	26
bread	3	razor	27
fire	4	mask	28
woman	5	shark	29
village	6	study	30
dinner	7	fierce	31
stop	8	temple	32
orange	9	military	33
lumber	10	gypsy	34
string	11	musician	35
knee	12	wrestle	36
forest	13	dwelling	37
sword	14	weapon	38
orchard	15	slumber	39
insect	16	admiral	40
grocer	17	medal	41
hatchet	18	arbor	42
merchant	19	garrison	43
anchor	20	dormitory	44
onion	21	chandelier	45
slipper	22	equestrian	46
arrow	23	pugelist	47
veil	24	rhythmic	48

Table 14

Original Item Parameters for Gates Word Recognition

	1956			1977		
ITEM #	a	b	c	a	b	c
1	0.344	-11.057	0.240	0.816	-4.700	0.250
2	0.099	-25.553	0.240	0.417	-6.080	0.250
3	1.176	-4.313	0.240	0.680	-5.693	0.250
4	0.739	-5.230	0.240	0.949	-4.578	0.250
5	0.620	-5.544	0.240	0.446	-7.437	0.250
6	0.511	-3.549	0.240	0.600	-2.433	0.250
7	0.912	-3.836	0.240	0.684	-4.853	0.250
8	0.810	-3.907	0.240	0.680	-4.516	0.250
9	0.698	-4.132	0.240	0.998	-3.888	0.250
10	1.359	-2.347	0.240	1.299	-2.470	0.250
11	1.004	-2.125	0.240	1.140	-2.859	0.250
12	1.002	-2.141	0.240	0.839	-3.072	0.250
13	0.788	-3.149	0.240	0.816	-3.286	0.250
14	1.148	-2.110	0.240	1.268	-2.368	0.250
15	1.324	-1.154	0.240	1.517	-1.282	0.250
16	1.282	-2.074	0.240	1.273	-2.662	0.250
17	1.144	-2.123	0.240	1.664	-1.474	0.250
18	0.741	-1.397	0.240	0.650	-1.544	0.250
19	0.909	-0.754	0.240	0.932	-0.861	0.250
20	0.935	-0.317	0.240	0.940	-1.426	0.250
21	1.120	-1.354	0.240	1.267	-2.020	0.250
22	1.337	-1.685	0.240	1.174	-2.534	0.250
23	1.328	-1.938	0.240	0.970	-2.897	0.250
24	0.744	-0.496	0.240	0.626	-0.152	0.250
25	1.619	-0.760	0.240	0.849	-1.200	0.250
26	1.229	-0.915	0.240	1.223	-1.808	0.250
27	1.326	-0.965	0.240	1.034	-1.429	0.250
28	0.450	-3.130	0.240	0.446	-3.471	0.250
29	1.023	-1.334	0.240	0.834	-3.874	0.250
30	1.676	-1.490	0.240	1.701	-2.209	0.250
31	1.993	-0.895	0.240	1.990	-0.923	0.250
32	1.158	-1.072	0.240	1.165	-1.582	0.250
33	0.893	-0.571	0.240	1.220	-0.870	0.250
34	2.000	0.525	0.249	1.990	-0.056	0.298
35	0.639	-1.207	0.240	0.987	-1.344	0.250
36	1.565	-0.726	0.240	1.990	-1.274	0.250
37	0.679	0.440	0.240	0.221	0.449	0.250
38	1.666	-0.548	0.260	1.229	-1.134	0.250
39	1.264	0.452	0.260	1.990	0.611	0.256

Table 14 (con't)

	1956			1977		
ITEM #	a	b	c	a	b	c
40	2.000	0.930	0.216	0.933	0.364	0.250
41	1.184	0.461	0.240	0.881	-0.086	0.250
42	1.896	0.882	0.250	0.894	0.925	0.203
43	2.000	1.224	0.145	1.990	1.519	0.158
44	0.788	0.322	0.240	0.685	-0.142	0.250
45	0.833	0.777	0.240	0.936	-1.054	0.250
46	0.521	2.654	0.240	0.315	3.894	0.250
47	0.072	10.005	0.000	1.990	3.983	0.250
48	0.719	1.617	0.240	1.990	1.592	0.250

Table 15

Rescaled Item Parameters for Gates Word Recognition

	1956			1977		
ITEM #	a	b	c	a	b	c
1		N/A			N/A	
2		N/A			N/A	
3		N/A			N/A	
4		N/A			N/A	
5		N/A			N/A	
6	0.511	-3.549	0.240	0.694	-1.896	0.240
7		N/A			N/A	
8		N/A			N/A	
9		N/A			N/A	
10	1.359	-2.347	0.240	1.502	-1.928	0.240
11	1.004	-2.125	0.240	1.318	-2.265	0.240
12	1.002	-2.141	0.240	0.970	-2.449	0.240
13	0.788	-3.149	0.240	0.944	-2.634	0.240
14	1.148	-2.110	0.240	1.466	-1.840	0.240
15	1.324	-1.154	0.240	1.754	-0.901	0.240
16	1.282	-2.074	0.240	1.471	-2.094	0.240
17	1.144	-2.123	0.240	1.924	-1.067	0.240
18	0.741	-1.397	0.240	0.752	-1.127	0.240
19	0.909	-0.754	0.240	1.077	-0.536	0.240
20	0.935	-0.317	0.240	1.087	-1.025	0.240
21	1.120	-1.354	0.240	1.464	-1.539	0.240
22	1.337	-1.685	0.240	1.357	-1.984	0.240
23	1.328	-1.938	0.240	1.121	-2.298	0.240
24	0.744	-0.496	0.240	0.724	0.077	0.240
25	1.619	-0.760	0.240	0.981	-0.830	0.240
26	1.229	-0.915	0.240	1.414	-1.356	0.240
27	1.326	-0.965	0.240	1.196	-1.028	0.240
28	0.450	-3.130	0.240	0.516	-2.794	0.240
29	1.023	-1.334	0.240	0.964	-3.142	0.240
30	1.676	-1.490	0.240	1.967	-1.703	0.240
31	1.993	-0.895	0.240	2.301	-0.590	0.240
32	1.158	-1.072	0.240	1.347	-1.160	0.240
33	0.893	-0.571	0.240	1.411	-0.544	0.240
34	2.000	0.525	0.249	2.301	0.160	0.249
35	0.639	-1.207	0.240	1.141	-0.954	0.240
36	1.565	-0.726	0.240	2.301	-0.894	0.240
37	0.679	0.440	0.240	0.256	0.597	0.240
38	1.666	-0.548	0.260	1.421	-0.773	0.260
39	1.264	0.452	0.260	2.301	0.737	0.260

Table 15 (con't)

	1956			1977		
ITEM #	a	b	c	a	b	c
40	2.000	0.930	0.216	1.079	0.524	0.216
41	1.184	0.461	0.240	1.018	0.134	0.240
42	1.896	0.882	0.250	1.034	1.008	0.250
43	2.000	1.224	0.145	2.301	1.522	0.145
44	0.788	0.322	0.240	0.791	0.086	0.240
45	0.833	0.777	0.240	1.082	-0.703	0.240
46	0.521	2.654	0.240	0.365	3.576	0.240
47		N/A			N/A	
48	0.719	1.617	0.240	2.301	1.585	0.240

N/A - This item was not included due to an extreme original difficulty parameter.

Table 16

The Number of Students Located at Each Level of Ability for the Gates Word Recognition Test

Ability	1956	1977*	Ability	1956	1977*
-3.00	6	1	0.20	225	307
-2.80	9	1	0.40	219	308
-2.60	8	1	0.60	257	431
-2.40	16	3	0.80	267	241
-2.20	29	5	1.00	210	197
-2.00	41	14	1.20	171	289
-1.80	51	19	1.40	113	156
-1.60	81	29	1.60	72	83
-1.40	109	48	1.80	39	90
-1.20	146	118	2.00	29	51
-1.00	179	146	2.20	16	50
-0.80	222	327	2.40	10	34
-0.60	259	286	2.60	11	11
-0.40	260	321	2.80	14	1
-0.20	227	310	3.00	1	1
0.00	205	455			

* The 1977 distribution following the rescaling procedure

Table 17

Aberrant Items at Specified Levels of Difference

Gates Word Recognition Test

Item #	Diff.	≥ 0.05	≥ 0.10	≥ 0.15
1	N/A			
2	N/A			
3	N/A			
4	N/A			
5	N/A			
6	0.0543	*		
7	N/A			
8	N/A			
9	N/A			
10	0.0139			
11	0.0213			
12	0.0127			
13	0.0053			
14	0.0078			
15	0.0299			
16	0.0070			
17	0.0723	*		
18	0.0327			
19	0.0314			
20	0.1409	*	*	
21	0.0396			
22	0.0210			
23	0.0096			
24	0.1038	*	*	
25	0.0456	*		
26	0.0749	*		
27	0.0085			
28	0.0048			
29	0.1005	*	*	
30	0.0238			
31	0.0609	*		
32	0.0248			
33	0.0490	*		
34	0.0943	*		
35	0.0498	*		
36	0.0518	*		
37	0.0838	*		
38	0.0395			
39	0.0899	*		
40	0.1209	*	*	

Table 17 (con't)

Item	Diff.	≥ 0.05	≥ 0.10	≥ 0.15
41	0.0759	*		
42	0.0534	*		
43	0.0551	*		
44	0.0454	*		
45	0.3016	*	*	*
46	0.0143			
47	N/A			
48	0.0945	*	*	

N/A - This item was not included due to an extreme original difficulty (b) parameter.

Table 18

Mean Student Abilities for the Gates Word Recognition Test

Level of aberrance		1956 students	1977 students	1956+1977 combined	# items removed	# items remaining
1.0	mean =	-0.167	0.224	0.050	0	48
	s.d. =	1.218	1.060	1.149		
0.15	mean =	-0.130	0.208	0.057	1	47
	s.d. =	1.240	1.060	1.156		
0.10	mean =	-0.033	0.263	0.131	6	42
	s.d. =	1.345	1.168	1.258		
0.05	mean =	0.037	0.276	0.170	21	27
	s.d. =	1.323	1.265	1.296		
	N =	3540	4430	7970		

test item are presented in Table 23. The results from the aberrant test items that were identified in Table 23 for each of the different aberrant levels were removed from the total test data. The selected test item data were then pooled across the two study years and reanalyzed by LOGIST. This was done in order to obtain new student abilities which did not include the results from the aberrant items. Mean abilities were then calculated for both the 1956 Grade III students and the 1977 Grade III students at each of the specified levels of aberrance and are presented in Table 24. The item characteristic curves for the California Mental Maturity Test are presented in Figures 89 to 147.

Table 19

Study Item Number Assignment for California Mental Maturity

Page #	Item #	Study Item #
3	1	1
3	2	2
3	3	3
3	4	4
3	5	5
3	6	6
3	7	7
3	8	8
3	9	9
3	10	10
4	1	11
4	2	12
4	3	13
4	4	14
4	5	15
4	6	16
4	7	17
4	8	18
4	9	19
4	10	20
4	11	21
4	12	22
5	1	23
5	2	24
5	3	25
5	4	26
5	5	27
5	6	28
5	7	29
5	8	30
5	9	31
5	10	32
5	11	33
5	12	34
6	1	35
6	2	36
6	3	37
6	4	38
6	5	39
6	6	40

Table 19 (con't)

Page #	Item #	Study Item #
6	7	41
6	8	42
6	9	43
6	10	44
6	11	45
6	12	46
7	1	47
7	2	48
7	3	49
7	4	50
7	5	51
7	6	52
7	1	53
7	2	54
7	3	55
7	4	56
7	5	57
7	6	58
8	1	59
8	2	60
8	3	61
8	4	62
8	5	63
8	6	64
8	7	65
8	8	66
8	9	67
8	10	68
8	11	69
8	12	70
9	1	71
9	2	72
9	3	73
9	4	74
9	5	75
9	6	76
9	7	77
9	8	78
9	9	79
9	10	80
9	11	81
9	12	82

Table 19 (con't)

Page #	Item #	Study Item #
9	13	83
9	14	84
9	15	85
9	16	86
9	17	87
9	18	88
9	19	89
9	20	90
9	21	91
9	22	92
9	23	93
9	24	94
9	25	95
9	26	96
9	27	97
9	28	98

Table 20

Original Item Paramaters for California Mental Maturity

	1956			1977		
ITEM #	a	b	c	a	b	c
1	1.961	0.050	0.210	1.965	0.227	0.343
2	1.961	0.028	0.189	1.965	0.213	0.338
3	0.758	-0.050	0.210	0.444	-0.048	0.275
4	1.961	0.121	0.183	1.965	0.294	0.328
5	0.091	-3.679	0.210	0.207	-1.069	0.275
6	0.010	-1.010	0.210	0.010	14.960	0.275
7	1.961	1.243	0.210	0.500	0.877	0.275
8	1.961	0.951	0.283	0.506	-0.309	0.275
9	1.961	1.133	0.179	0.517	-0.292	0.275
10	1.961	1.173	0.234	0.488	0.012	0.275
11	0.790	-3.491	0.210	0.889	-3.538	0.275
12	1.405	-2.824	0.210	1.312	-2.852	0.275
13	1.422	-3.014	0.210	1.212	-3.050	0.275
14	1.275	-2.890	0.210	1.485	-3.008	0.275
15	1.200	-3.084	0.210	1.026	-2.938	0.275
16	0.740	-3.098	0.210	1.404	-2.785	0.275
17	1.150	-2.705	0.210	0.967	-3.019	0.275
18	0.836	-2.918	0.210	1.050	-2.977	0.275
19	0.403	-2.363	0.210	0.613	-2.449	0.275
20	0.387	-2.597	0.210	0.611	-2.570	0.275
21	0.372	-3.150	0.210	0.558	-2.796	0.275
22	0.452	3.672	0.239	0.740	1.375	0.275
23	0.753	-3.875	0.210	0.856	-3.527	0.275
24	1.559	-3.098	0.210	1.159	-3.523	0.275
25	1.121	-3.322	0.210	1.024	-3.213	0.275
26	0.481	-3.415	0.210	0.395	-4.648	0.275
27	1.219	-3.174	0.210	1.088	-3.428	0.275
28	0.559	-2.878	0.210	0.873	-2.750	0.275
29	0.804	-3.082	0.210	0.832	-3.232	0.275
30	0.824	-3.017	0.210	0.845	-3.470	0.275
31	0.788	-2.614	0.210	0.772	-3.151	0.275
32	0.775	-3.010	0.210	0.708	-2.918	0.275
33	0.614	-2.889	0.210	0.703	-3.401	0.275
34	0.646	-3.202	0.210	0.453	-4.267	0.275
35	0.376	-7.132	0.210	0.592	-4.317	0.275
36	0.294	-3.914	0.210	0.395	-3.503	0.275
37	0.269	-3.765	0.210	0.350	-5.460	0.275
38	1.961	2.115	0.232	0.367	4.588	0.275
39	0.247	-1.287	0.210	0.355	-0.490	0.275

Table 20 (con't)

	1956			1977		
ITEM #	a	b	c	a	b	c
40	0.096	-2.142	0.210	0.235	-1.102	0.275
41	0.250	-1.347	0.210	0.529	-0.598	0.275
42	0.943	2.823	0.293	0.789	1.885	0.275
43	0.127	-5.242	0.210	0.298	-3.291	0.275
44	0.141	-6.240	0.210	0.299	-3.046	0.275
45	0.267	-0.667	0.210	0.567	-0.758	0.275
46	0.141	0.129	0.210	0.299	-0.859	0.275
47	0.151	-7.067	0.210	0.254	-4.621	0.275
48	0.351	-5.504	0.210	0.289	-6.150	0.275
49	0.555	-5.698	0.210	0.683	-4.076	0.275
50	0.121	-6.428	0.210	0.288	-4.296	0.275
51	0.122	-4.668	0.210	0.219	-2.105	0.275
52	0.249	-6.874	0.210	0.508	-3.381	0.275
53	0.356	-0.617	0.210	0.553	-1.438	0.275
54	0.469	0.200	0.210	1.363	-0.017	0.403
55	0.810	2.089	0.295	1.206	0.768	0.327
56	1.961	1.755	0.101	1.965	0.896	0.137
57	1.961	1.851	0.182	1.965	0.740	0.246
58	0.692	4.012	0.142	0.670	2.932	0.140
59	0.470	-3.852	0.210	0.611	-3.109	0.275
60	0.164	-6.059	0.210	0.260	-4.646	0.275
61	0.403	-5.323	0.210	0.587	-3.834	0.275
62	0.204	-3.681	0.210	0.459	-1.848	0.275
63	0.393	-1.773	0.210	0.505	-1.255	0.275
64	0.388	-3.953	0.210	0.581	-2.471	0.275
65	1.855	1.867	0.091	1.965	1.891	0.062
66	0.024	-24.997	0.210	0.130	-5.628	0.275
67	0.133	-7.350	0.210	0.403	-2.955	0.275
68	0.221	-6.228	0.210	0.494	-2.797	0.275
69	0.486	0.688	0.210	1.041	0.644	0.275
70	0.592	1.887	0.248	1.577	1.171	0.275
71	0.541	-6.398	0.210	0.807	-4.182	0.275
72	0.612	-5.161	0.210	0.818	-4.315	0.275
73	0.480	-4.937	0.210	0.297	-3.685	0.275
74	0.205	-9.265	0.210	0.335	-4.383	0.275
75	0.982	-4.311	0.210	0.807	-3.937	0.275
76	0.397	-5.210	0.210	0.420	-4.911	0.275
77	0.197	-7.666	0.210	0.301	-3.712	0.275
78	0.527	-4.870	0.210	0.718	-3.602	0.275
79	0.505	-5.552	0.210	0.954	-3.616	0.275
80	0.565	-3.672	0.210	0.754	-2.581	0.275
81	0.779	-4.133	0.210	0.950	-3.189	0.275

Table 20 (con't)

	1956			1977		
ITEM #	a	b	c	a	b	c
82	0.633	-5.306	0.210	1.040	-3.073	0.275
83	0.463	-5.353	0.210	0.789	-3.790	0.275
84	0.751	-4.276	0.210	0.946	-3.183	0.275
85	0.773	-4.315	0.210	1.040	-3.100	0.275
86	0.685	-5.013	0.210	0.753	-4.133	0.275
87	0.111	3.091	0.210	0.294	0.556	0.275
88	0.604	-4.632	0.210	0.964	-3.231	0.275
89	0.256	-3.136	0.210	0.670	-2.861	0.275
90	0.255	-0.733	0.210	0.628	-2.107	0.275
91	0.395	-5.232	0.210	0.791	-3.125	0.275
92	0.302	-4.187	0.210	0.292	-3.937	0.275
93	0.208	-0.218	0.210	0.253	-0.319	0.275
94	0.411	3.810	0.087	0.464	2.054	0.096
95	0.270	-1.878	0.210	0.302	-1.078	0.275
96	0.830	2.155	0.291	0.387	0.591	0.275
97	0.071	1.468	0.210	0.227	0.075	0.275
98	0.074	-11.413	0.210	0.184	-4.283	0.275

Table 21

Rescaled Item Parameters for California Mental Maturity

	1956			1977		
ITEM #	a	b	c	a	b	c
1	1.961	0.050	0.210	1.495	0.901	0.210
2	1.961	0.028	0.189	1.495	0.883	0.189
3	0.758	-0.050	0.210	0.338	0.540	0.210
4	1.961	0.121	0.183	1.495	0.990	0.183
5	0.091	-3.679	0.210	0.158	-0.802	0.210
6		N/A			N/A	
7	1.961	1.243	0.210	0.380	1.756	0.210
8	1.961	0.951	0.283	0.385	0.197	0.283
9	1.961	1.133	0.179	0.393	0.219	0.179
10	1.961	1.173	0.234	0.371	0.618	0.234
11	0.790	-3.491	0.210	0.676	-4.047	0.210
12	1.405	-2.824	0.210	0.998	-3.146	0.210
13	1.422	-3.014	0.210	0.922	-3.405	0.210
14	1.275	-2.890	0.210	1.130	-3.351	0.210
15	1.200	-3.084	0.210	0.781	-3.258	0.210
16	0.740	-3.098	0.210	1.068	-3.057	0.210
17	1.150	-2.705	0.210	0.736	-3.364	0.210
18	0.836	-2.918	0.210	0.799	-3.309	0.210
19	0.403	-2.363	0.210	0.466	-2.616	0.210
20	0.387	-2.597	0.210	0.465	-2.775	0.210
21	0.372	-3.150	0.210	0.425	-3.072	0.210
22	0.452	3.672	0.239	0.563	2.411	0.239
23	0.753	-3.875	0.210	0.651	-4.033	0.210
24	1.559	-3.098	0.210	0.882	-3.027	0.210
25	1.121	-3.322	0.210	0.779	-3.620	0.210
26		N/A			N/A	
27	1.219	-3.174	0.210	0.828	-3.903	0.210
28	0.559	-2.878	0.210	0.665	-3.012	0.210
29	0.804	-3.082	0.210	0.633	-3.644	0.210
30	0.824	-3.017	0.210	0.643	-3.957	0.210
31	0.788	-2.614	0.210	0.587	-3.538	0.210
32	0.775	-3.010	0.210	0.538	-3.232	0.210
33	0.614	-2.889	0.210	0.535	-3.868	0.210
34		N/A			N/A	
35		N/A			N/A	
36	0.294	-3.914	0.210	0.301	-4.001	0.210
37		N/A			N/A	
38		N/A			N/A	
39	0.247	-1.287	0.210	0.270	-0.041	0.210

Table 21 (con't)

	1956			1977		
ITEM #	a	b	c	a	b	c
40	0.096	-2.142	0.210	0.179	-0.845	0.210
41	0.250	-1.347	0.210	0.402	-0.182	0.210
42	0.943	2.823	0.293	0.600	3.081	0.293
43		N/A			N/A	
44		N/A			N/A	
45	0.267	-0.667	0.210	0.431	-0.393	0.210
46	0.141	0.129	0.210	0.227	-0.526	0.210
47		N/A			N/A	
48		N/A			N/A	
49		N/A			N/A	
50		N/A			N/A	
51		N/A			N/A	
52		N/A			N/A	
53	0.356	-0.617	0.210	0.421	-1.287	0.210
54	0.469	0.200	0.210	1.037	0.580	0.210
55	0.810	2.089	0.295	0.918	1.613	0.295
56	1.961	1.755	0.101	1.495	1.781	0.101
57	1.961	1.851	0.182	1.495	1.576	0.182
58		N/A			N/A	
59	0.470	-3.852	0.210	0.465	-3.483	0.210
60		N/A			N/A	
61		N/A			N/A	
62	0.204	-3.681	0.210	0.349	-1.825	0.210
63	0.393	-1.773	0.210	0.384	-1.047	0.210
64	0.388	-3.953	0.210	0.442	-2.645	0.210
65	1.855	1.867	0.091	1.495	3.089	0.091
66		N/A			N/A	
67		N/A			N/A	
68		N/A			N/A	
69	0.486	0.688	0.210	0.792	1.449	0.210
70	0.592	1.887	0.248	1.199	2.143	0.248
71		N/A			N/A	
72		N/A			N/A	
73		N/A			N/A	
74		N/A			N/A	
75		N/A			N/A	
76		N/A			N/A	
77		N/A			N/A	
78		N/A			N/A	
79		N/A			N/A	
80	0.565	-3.672	0.210	0.573	-2.789	0.210
81		N/A			N/A	

TABLE 20 (CON'T)

	1956			1977		
ITEM #	a	b	c	a	b	c
82		N/A			N/A	
83		N/A			N/A	
84		N/A			N/A	
85		N/A	N/A			
86		N/A			N/A	
87	0.111	3.091	0.210	0.224	1.333	0.210
88		N/A			N/A	
89	0.256	-3.136	0.210	0.510	-3.157	0.210
90	0.255	-0.733	0.210	0.478	-2.166	0.210
91		N/A			N/A	
92		N/A			N/A	
93	0.208	-0.218	0.210	0.192	0.184	0.210
94	0.411	3.810	0.087	0.353	3.303	0.087
95	0.270	-1.878	0.210	0.230	-0.814	0.210
96	0.830	2.155	0.291	0.294	1.380	0.291
97	0.071	1.468	0.210	0.173	0.702	0.210
98		N/A			N/A	

N/A - This item was not included due to an extreme original difficulty parameter.

Table 22

The Number of Students Located at Each Level of Ability for the California Mental Maturity Test

Ability	1956	1977*	Ability	1956	1977*
-3.00	7	8	0.20	223	267
-2.80	19	25	0.40	367	136
-2.60	28	23	0.60	381	303
-2.40	32	23	0.80	229	133
-2.20	46	23	1.00	194	340
-2.00	43	75	1.20	172	207
-1.80	64	79	1.40	123	383
-1.60	74	55	1.60	64	160
-1.40	105	55	1.80	34	291
-1.20	96	148	2.00	14	131
-1.00	115	171	2.20	9	179
-0.80	162	93	2.40	4	129
-0.60	213	93	2.60	3	44
-0.40	263	190	2.80	3	44
-0.20	205	196	3.00	2	51
0.00	160	131			

* The 1977 distribution following the rescaling procedure

Table 23

Aberrant Items at Specified Levels of Difference

California Mental Maturity

Item #	Diff.	≥ 0.05	≥ 0.10	≥ 0.15
1	0.1948	*	*	*
2	0.2008	*	*	*
3	0.1190	*	*	
4	0.2051	*	*	*
5	0.0600	*		
6	N/A			
7	0.1531	*	*	*
8	0.1921	*	*	*
9	0.2364	*	*	*
10	0.2014	*	*	*
11	0.0035			
12	0.0048			
13	0.0051			
14	0.0058			
15	0.0110			
16	0.0147			
17	0.0095			
18	0.0087			
19	0.0359			
20	0.0351			
21	0.0140			
22	0.0537	*		
23	0.0034			
24	0.0039			
25	0.0061			
26	N/A			
27	0.0038			
28	0.0208			
29	0.0056			
30	0.0096			
31	0.0143			
32	0.0200			
33	0.0212			
34	N/A			
35	N/A			
36	0.0068			
37	N/A			
38	N/A			
39	0.0912	*		
40	0.0262			

Table 23 (con't)

Item #	Diff.	≥ 0.05	≥ 0.10	≥ 0.15
41	0.0707	*		
42	0.0249			
43	N/A			
44	N/A			
45	0.0432			
46	0.0540	*		
47	N/A			
48	N/A			
49	N/A			
50	N/A			
51	N/A			
52	N/A			
53	0.0834	*		
54	0.1042	*	*	
55	0.0443			
56	0.0195			
57	0.0542	*		
58	N/A			
59	0.0133			
60	N/A			
61	N/A			
62	0.0366			
63	0.0688	*		
64	0.0428			
65	0.0910	*		
66	N/A			
67	N/A			
68	N/A			
69	0.1404	*	*	
70	0.0921	*		
71	N/A			
72	N/A			
73	N/A			
74	N/A			
75	N/A			
76	N/A			
77	N/A			
78	N/A			
79	N/A			
80	0.0312			
81	N/A			
82	N/A			

Table 23 (con't)

Item #	Diff.	≥ 0.05	≥ 0.10	≥ 0.15
83	N/A			
84	N/A			
85	N/A			
86	N/A			
87	0.0411			
88	N/A			
89	0.0967	*	*	
90	0.1903	*	*	*
91	N/A			
92	N/A			
93	0.0274			
94	0.0573	*		
95	0.0918	*		
96	0.1705	*	*	*
97	0.0326			
98	N/A			

N/A - This item was not included due to an extreme original difficulty (b) parameter.

Table 24

Mean Student Abilities for the California Mental Maturity

Level of aberrance		1956 students	1977 students	1956+1977 combined	# items removed	# items remaining
1.0	mean =	-0.320	0.233	-0.010	0	98
	s.d. =	0.927	1.074	1.049		
0.15	mean =	-0.266	0.235	0.014	9	89
	s.d. =	0.885	1.099	1.040		
0.10	mean =	-0.264	0.241	0.019	13	85
	s.d. =	0.862	1.112	1.040		
0.05	mean =	-0.243	0.229	0.021	25	73
	s.d. =	0.879	1.112	1.043		
	N =	3443	4378	7821		

V. DISCUSSION AND CONCLUSIONS

The discussion will be presented in the following manner:

1. an introduction that will discuss in general terms the way in which the results were reviewed;
2. a discussion of the results of each of the specific tests;
3. a comparison and discussion of the implications of the item statistics that were computed by latent trait theory methodology verses the implications of the item statistics that were computed by classical test theory methodology; and
4. a concluding summary.

A. Introduction

Each item on each test falls into one of the following categories:

1. The item was very easy for both groups of students. Hence, there is no difference in the pattern of student response between the two study years.
2. The item had similar item characteristics curves for both study years. This would indicate that students having a similar ability in either 1956 or 1977 had an equal probability of getting this item correct. Therefore, this item was not biasing the results toward

either group of students for any one study year.

3. The item produced different item characteristic curves for each study year. This would indicate that persons having similar ability levels in 1956 and 1977 may not have had the same probability of getting the item correct. For some reason, then, this item would appear to be measuring a different trait(s) over all student abilities or some portion of the student ability continuum. A condition of aberrance then exists for this item. The degree of aberrance would indicate the degree of item bias that exists. Depending on the conditions that prevail, the bias may be toward either the 1956 student or the 1977 student.

4. There are four ways that the item parameters obtained from the 1977 students may differ from the item parameters obtained from the 1956 students.
    a. The discrimination (a) parameter is less in 1977 than it was in 1956 and the difficulty (b) parameter is less as well.
    b. The discrimination (a) parameter is less in 1977 than it was in 1956, but the difficulty (b) parameter is higher.
    c. The discrimination (a) parameter is higher in 1977 than it was in 1956 and the difficulty (b) parameter is lower.
    d. The discrimination (a) parameter is higher in 1977 than it was in 1956 and the difficulty (b)

parameter is higher as well.

The aberrant items will be discussed in terms of these four categories in order to determine if there was any consistent bias pattern among the aberrant items.

In order to ascertain the degree of item aberrance, it was necessary to have a measure of the reliability of the item characteristic curve. The true error of this curve can be established by determining the probability of success for each value of ability as predicted by the model and then using the data to assess the actual degree of success for each student at each ability level for each item. Because of the expense of approaching the problem of standard error in this manner, it was considered that a pseudo standard error could be calculated in the following manner:

$$\hat{s}_{S.E.}(\theta) = \frac{P_i(\theta) \cdot Q_i(\theta)}{n_i(\theta)} \qquad (8)$$

where:

$\hat{s}_{S.E.}(\theta)$ = the estimated standard error of the probability of success for ability level θ on a given test item;

$Q_i(\theta)$ = the probability of an incorrect answer for ability level θ on item i;

$n_i(\theta)$ = the number of examinees having the ability level θ.

The estimated standard error is indicated on each Figure: depicting the relationship between the item characteristic curve, obtained from each study year for each item, and the estimated standard error. This is included only as a rough guideline to standard error and is not intended to represent the actual error between the model and the data.

B. Discussion of Individual Test Results

Raven Progressive Matrices Test

Items 1, 2, 3, 4, 5, 6, 13, 14, 15, 25, and 26 were identified in Table 2 as having a difficulty (b) parameter that was less than -4.00 as calculated from either the 1956 or the 1977 data. The item characteristic curves for these items would take the form of Figure 1. All of these items were found by the students to be so easy that there was little or no discrimination between the high ability students and the low ability students. This means that nearly all the students from both study years answered these items correctly. This result clearly reflects Raven's test format. Items 1 to 12 are from Raven's subtest A which is the easiest subtest within this test. Items 13 to 24 are from Raven's subtest Ab, which was to be slightly harder than subtest A. Items 25 to 36 are from Raven's subtest B; the hardest of the three subtests. The first few items in each subtest were found to be easy by the students of both study years and, consequently, these items demonstrate no bias toward either group of students.

The characteristic curves for items 7, 9, 10, 17, 19, 22, 27, 28, and 29 (Figures 2,4,5,9,11,14,17,18, & 19 respectively) indicate that these items performed in a similar manner when administered to both study years; i.e., a student of a given ability level had an equal probability of answering one of these items correctly, whether or not he belonged to the 1956 group of students or to the 1977 group of students (see Table 5).

The item characteristic curves (Figures 1 to 26) in conjunction with the results presented in tables 5 and 25, form the basis of the following discussion of the location of where the achievement differences between the 1956 and the 1977 students occurred when the examinee ability continuum was roughly divided into three parts: low ($\theta < -1$), medium ($-1 < \theta > +1$), and high ($\theta > +1$).

Table 5 indicates that items 12, 30, 34, and 35 had an estimated item difference of at least 0.10. Figure 20 shows that for item 30, the probability of answering this item correctly was higher for the low- and medium-ability 1977 students than for the 1956 students of similar ability.

Figure 24 shows that for item 34, the medium-ability students in 1977 had a higher probability of answering this item correctly than did the students of similar ability in 1956. The low- and high-ability students in 1977 had the same probability of success as their ability counterpart

Table 25

Location of Item Bias by Ability Level for Raven's Progressive Matrices Test

	Ability Level					
	1956			1977		
item #	low	med.	high	low	med.	high
1						
2						
3						
4						
5						
6						
7						
8	*	*	*			
9						
10						
11		*	*			
12		*				
13						
14						
15						
16				*		
17						
18		*				
19						
20		*				
21				*		
22						
23					*	
24			*			
25						
26						
27						
28						
29						

Table 25 (con't)

Ability Level

item #	1956 low	1956 med.	1956 high	1977 low	1977 med.	1977 high
30				*	*	
31			*	*		
32					*	
33					*	
34					*	
35					*	
36			*			

* Indicates those areas in the ability continuum where the students from a given study year had a higher probability of getting a given item correct than did the students of a similar ability level from the opposing study year.

1956 students. A similar pattern to that reported for item 34 (Figure 24) is visible for item 35 (Figure 25). Figure 7 shows that for item 12 the 1956 medium-ability students had a higher probability of answering this item correctly than did the 1977 medium-ability student.

Table 5 shows that items 8, 11, 16, 18, 20, 21, 23, 24, 31, 32, 33, and 36 had an estimated item difference of at least 0.05. There does not appear to be any single pattern of difference among these items between the time they were first administered in 1956 and the time when they were readministered in 1977.

Figure 8 (item 16) indicates that low-ability students in 1977 had a higher probability of answering the item correctly than did the low-ability 1956 students. The higher ability students of both study years appeared to have an equal probability of answering this item correctly.

Figures 10 and 12 (items 18 and 20) show a similar pattern where the medium-ability students in 1956 had a slightly higher probability of answering these two items correctly than did the 1977 students. This also appears true for item 11 (Figure 6) except that the high ability subjects may be affected as well. The reverse is true for item 23 (Figure 15); i.e., the medium-ability students in 1977 have a higher probability of answering this item correctly than did the 1956 medium-ability students. However, the pseudo standard error pattern shown in Figures 6 and 15 may indicate that even though there is an estimated item difference greater than 0.05, this difference is distributed across a wide range of abilities and may not, in fact, be significantly biasing any student located at any specific ability level in either study year.

Figure 3 (item 8) shows that the 1956 students had a higher probability of answering this item correctly than did the 1977 students over all levels of abilities. The ability for this item to discriminate well among the students of different ability levels for either study group is, however, quite low as is indicated by the overall flat curve. Because of this, the usefulness of this item could be questioned.

Figures 13, 22, and 23 indicate that for items 21, 32, and 33, the medium-ability 1977 students had a higher probability of answering these items correctly than did the 1956 students.

Figures 16 and 26 indicate that for items 24 and 36, the high-ability 1956 students had a higher probability of answering this item correctly than did the 1977 students.

Figure 21 indicates that for item 31 the low-ability 1956 students had a lower probability of answering this item correctly than did the 1977 students and, further to this, the high-ability 1956 students had a higher probability of answering this item correctly.

The different achievement means were reported in Table 6. If all 36 items are included in the students' evaluation, then the mean ability difference between the 1956 students and the 1977 students is 0.48 ability units or roughly a half a standard deviation unit on the ability scale in favor of the 1977 students. This difference indicates that the 1977 students were able to answer Raven's test items more proficiently than did the 1956 students.

However, Table 5 indicated that when the Raven Progressive Matrices test was administered, not all the test items were measuring the underlying latent trait(s) in the same manner for each study group.

When the data from the items that had an estimated item difference of 0.10 or greater were removed before the mean achievement difference score was calculated, the difference between the two study groups went down to 0.42 ability units in favor of the 1977 group. The decrease of the mean achievement difference from 0.48 to 0.42 ability units would indicate that those items that were removed were slightly favoring the 1977 group of students.

When the data from the items that had an estimated item difference of 0.05 or greater were removed prior to the calculation of the mean achievement difference between the two study groups, this score equalled 0.46 ability units in favor of the 1977 students. The fact that the mean achievement difference score has remained almost constant across all conditions of aberrant item removal would suggest that some of the aberrant items in the test favored the 1956 students while others favored the 1977 students. By reviewing the item characteristic curves (Figures 2 to 26), it can be observed that there is no consistent pattern of aberrance in favor of either the 1956 group or the 1977 group. Items 8, 11, 12, 18, 20, 24, 31, and 36 indicate that the 1956 students had a higher probability of getting these items correct. The 1977 students had a higher probability than the 1956 students of getting items 16, 21, 23, 30, 32, 33, 34, and 35 correct.

Table 26 presents a summary of how the 1977 item parameters differ in relation to the 1956 item parameters for each of the aberrant items. A chi square of 0.41 (df=1) further verifies that there was no consistent item parameter difference pattern between the two study groups.

The mean ability difference of 0.46 would further tend to indicate that there were fewer 1977 low-ability students. Thus, the probable difference of scores for any given item may not have a significant impact if it is favoring the 1956 low ability students.

The mean achievement difference of approximately 0.5 of a standard ability unit, becomes even more significant when the mean ages of the two study groups are taken into account. The mean age of the 1956 students reported by Clarke et al. (1977, p. 10) was 108.9 months (N = 3,843, S.D. = 7.37), while the mean age of the 1977 students was 106.37 months (N = 4,720, S.D. = 5.59). Raven (1965) reports that as a child gets older he is expected to perform better on the Progressive Matrices test. Thus, the fact that the 1977 students had a higher mean ability than did the 1956 students at an earlier age indicates that the 1977 students,

Table 26

The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for Raven's Progressive Matrices

		Discrimination (a)	
		1977 < 1956	1977 > 1956
Difficulty (b)	1977 < 1956	16, 21, 30, 32, 33, 34, 35	23
	1977 > 1956	8, 18, 20, 24, 31, 36	11, 12

on the average, are more intelligent than the 1956 students.[7]

Gates Paragraph Reading Test

Items 1, 2, 3, 4, 5, 6, 9, 11, and 18 were identified from Table 8 as having a difficulty (b) parameter that was less than -4.00 as calculated from either the 1956 or the 1977 data. The item characteristic curves for these items would take the form of Figure 1. The students from both study years found these items to be easy. Consequently, these items were not found to be biased toward either the 1956 students or the 1977 students.

The item characteristic curves (Figures 27 to 49) in conjunction with the results presented in tables 11 and 27, form the basis of the following discussion of the location of where the achievement differences between the 1956 and the 1977 students occurred when the examinee ability continuum was roughly divided up into three parts: low ($\theta < -1$), medium ($-1 < \theta > +1$), and high ($\theta > +1$).

Items 7, 10, 12, 13, 14, 15, 16, 17, 21, 25, 26, 28, 29, 31, and 32 (Figures 27, 29, 30, 31, 32, 33, 34, 35, 38, 42, 43, 45, 46, 48, and 49, respectively) had estimated item differences of less than 0.05. These was no evidence of item bias with these items and students of a given ability level in 1956 or 1977 would appear to have an equal probability of answering these items correctly.

---

[7] Accepting the premise that Raven's Progressive Matrices test does, in fact, measure intelligence.

Table 27

Location of Item Bias by Ability Level for

Gate's Paragraph Reading Test

	Ability Level					
	1956			1977		
item #	low	med.	high	low	med.	high
1						
2						
3						
4						
5						
6						
7						
8				*	*	*
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19		*				
20		*				
21						
22				*		
23		*				
24		*				
25						
26						
27		*				
28						
29						
30			*			
31						
32						

* Indicates those areas in the ability continuum where the students from a given study year had a higher probability of getting a given item correct than did the students of a similar ability level from the opposing study year.

There was only one item that had an estimated item difference of greater than 0.15. This was item number 8 (Figure 28). While the 1977 students appeared to have a higher probability of answering this item correctly than did the 1956 students of similar ability levels, there appeared to be some confusion over this item by students of both study years. This is evidenced by the relatively high value of the lower asymptote and the shallow slope of the item characteristic curves. Item number 8 asks, "Which tree does not lose its leaves, " and shows a maple tree, an elm tree, and a pine tree. The confusion appears to result from whether or not the "needles" on the pine tree constitute "leaves". While this makes the item a poor one, it is still interesting to note that the 1977 students had a higher probability of answering this item correctly.

Items 19 and 23 (Figures 36 and 40, respectively) had estimated item difference greater than 0.10 along with the previously mentioned item number 8. For both of these items, the medium-ability 1956 student had a higher probability of answering this item correctly. There are no obvious reasons why these items would be biased in favor of the 1956 students. Although, one could speculate that the fashions in clothes in 1956 verses 1977 and the social studies curriculum changes over 21 years, both contribute to the observed student response pattern.

In addition to those items previously reported, items 20, 22, 24, 27, and 30 (Figures 37, 39, 41, 44, and 47, respectively) had estimated item differences greater than 0.05 (see Table 11). The 1956 students appear to have a higher probability of answering all these items correctly except for item 22 where the low ability 1977 students had a higher probability of answering this item correctly than did the low-ability students in 1956. The results obtained from the 1977 students on item 22 may be related to the the fact that more 1977 students answered the item correctly than did the 1956 students and, consequently, the model fit may be invalid due to lack of data for the extreme values of ability or, that the instruction "Draw a line around" was more confusing to the low-ability 1956 students than the low-ability 1977 students. The bias of the remaining items may be attributed to the use of term "modern" in the test directions and to the fact that the pictures incorporated in the test item would not appear "modern" to the 1977 students. Another cause of the apparent bias can be seen in the relationship between the 1956 students' environment and the environment of the 1977 students.

Table 12 indicated that when the mean achievement difference was calculated using all the test items, the 1977 students performed better than the 1956 students by 0.154 ability units. When the data from the aberrant items were removed prior to the computation of the mean achievement difference, the following pattern appears.

Recall that the 1977 students had a higher probability of getting item 8 correct over all ability levels than did the 1956 students of the same ability level. When the data from this item were removed and the mean achievement difference was recalculated, the 1977 students performed better than the 1956 students by only 0.139 ability units. The drop in the mean achievement difference would substantiate that this item was, in fact, favoring the 1977 students and, consequently, the data from this item should be removed before making any comparisons between the study groups.

There were, however, other items that appeared to favor the 1956 students, and when the data from these items (19 and 23) were removed along with the data from item 8, the mean achievement difference rose to an ability difference of .182 in favor of the 1977 students. The data for those items having an estimated item difference of 0.05 or greater produced a mean achievement difference of 0.189 in favor of the 1977 students.

Table 28 presents a summary of the 1977 item parameters in relation to the 1956 item parameters for each of the aberrant items. It is interesting to note that all the aberrant items had better discrimination properities in 1956 than in 1977. This could be a result of the test being out of date and the 1977 students found these items more confusing than did the 1956 students. Two of the aberrant items (8 & 22) were found to be easier for the 1977 students, while the remaining aberrant items favored the 1956 students.

Table 28

The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for Gates Paragraph Reading

		Discrimination (a)	
		1977 < 1956	1977 > 1977
Difficulty (b)	1977 < 1956	8, 22	
	1977 > 1956	19, 20, 23, 24, 27, 30	

These results would tend to indicate that the 1977 students performed better than would be initially considered when assessing the mean achievement diffference that incorporated all test items. Those items that appeared to favor the 1956 students included items dealing with "modern" fans, the act of sewing, a train schedule, and a "modern" use of flax (oil based paints). As the 1977 student interacts with air conditioning, ready-made clothes, planes, and latex or acrylic based paints as opposed to fans, sewing, trains, and oil base paints. It is understandable that the 1977 students may not have the same probability of answering those items correctly than did the 1956 students.

If the age difference previously reported is again taken into consideration, it would appear that the average 1977 student has more ability to read and understand a statement directing him/her to perform a particular task than did the average 1956 student.

Gates Word Recognition Test

Items 1, 2, 3, 4, 5, 7, 8, 9, and 47 were identified from Table 14 as having a difficulty (b) parameter that was less than -4.00 or greater than +4.00, as calculated from either the 1956 or 1977 data. The item characteristic curves for those items that had a difficulty (b) parameter of less than -4.00 would take the form of Figure 1. The item characteristic curves for those items that had a difficulty parameter of greater than +4.00 would take the form of Figure 148. These items were not found biasing either the

1956 students or the 1977 students because these items were either so easy or so difficult, that the students from both study years either got them correct in the case of very easy items, or wrong, in the case of very difficult items.

Items 10, 11, 12, 13, 14, 15, 16, 18, 19, 21, 22, 23, 27, 28, 30, 32, 38, and 46 (Figures 51, 52, 53, 54, 55, 56, 57, 59, 60, 62, 63, 64, 68, 69, 71, 73, 79, and 87, respectively) had estimated item differences of less than 0.05. There was no evidence of item bias with these items; i.e., examinees of a given ability level in either 1956 or 1977 would appear to have an equal probability of answering these items correctly.

The item characteristic curves (Figures 50 to 88) in conjunction with the results presented in Tables 17 and 29, form the basis of the following discussion of where the location of the achievement differences between the 1956 and the 1977 students occurred when the examinee ability continuum was roughly divided up into three parts: low ($\theta < -1$), medium ($-1 < \theta > +1$), and high ($\theta > +1$).

Item 45 (Figure 86, chandelier) had an estimated item difference of greater than 0.15. The 1977 medium and high ability students had a higher probability of answering this item correctly than did the 1956 students of similar ability levels. The lower asymptote would indicate that guessing was high for the lower ability students of both study years. The word "chandelier" appears to be a word that was known by both the medium and high ability 1977 students, but only the

Table 29

Location of Item Bias by Ability Level for Gates Word Recognition Test

	Ability Level					
	1956			1977		
item #	low	med.	high	low	med.	high
1						
2						
3						
4						
5						
6	*	*				
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17	*	*				
18						
19						
20					*	
21						
22						
23						
24		*				
25		*				
26				*	*	
27						
28						
29				*	*	

Table 29 (con't)

	Ability Level					
	1956			1977		
item #	low	med.	high	low	med.	high
30						
31	*	*				
32						
33	*				*	
34					*	
35	*				*	
36					*	
37			*	*		
38						
39		*				
40					*	
41					*	
42			*		*	
43			*			
44					*	
45					*	*
46						
47						
48		*				*

* Indicates those areas in the ability continuum where the students from a given study year had a higher probability of getting a given item correct than did the students of a similar ability level from the opposing study year.

very high ability 1956 students.

Items 20, 24, 29, 40, and 48 (Figures 61, 65, 70, 81, and 88, respectively) had an estimated item difference greater than 0.10 (along with the previously mentioned item 45. The results for item 29 (shark) indicated that the 1977 low and medium ability students had a higher probability of getting this item correct than did the 1956 students of similar ability. The recent exposure that the 1977 students had to films depicting sharks (e.g., "Jaws") could be an example of how material external to the school environment can influence a student's learning.

Items 20 and 40 (Figures 61 and 81) indicate that the 1977 medium ability students had a higher probability of understanding the meaning of admiral than did the 1956 medium ability students. A greater number of students from both study years correctly answered the anchor (item 20) question than the admiral question (item 40). This may be due to the unfamiliar way in which the admiral was drawn, but the item did not favor the 1956 group of students, but rather the 1977 group of students. Item 24 (Figure 65, veil) favored the 1956 medium ability students. Veils were more part of the fashions in the fifties than the seventies, so it is reasonable that the 1956 students would have a higher probability of getting this item correct. Item 48 (Figure 88, rhythmic) shows the 1956 medium ability students and the 1977 high ability students as having a higher probability of getting this item correct than did the students of similar ability in the opposing group of students. Overall, this item was found to be difficult for both groups of students, but the 1977 students in general would appear to have a slightly higher probability of getting this item correct.

Items 6, 17, 25, 26, 31, 33, 34, 35, 36, 37, 39, 41, 42, 43, and 44 (Figures 50, 58, 66, 67, 72, 74, 75, 76, 77, 78, 80, 82, 83, 84, and 85 respectively), in addition to those items previously reported, had estimated item differences of greater than 0.05 (see Table 17). The item characteristic curve plots show a variety of differences between the different ability groups, between the two study groups.

Items 6, 17, and 31 (Figures 50, 58, and 72; village, grocer, and fierce) indicate that the 1956 low and medium ability students had a higher probability of getting these items correct than the 1977 students of similar ability. The 1977 student would probably refer to the store manager or the produce manager rather than the grocer. The use of the term village in relation to the picture associated with item 6, has been superceded by the word 'town' and the picture of a large cat in item 31 was not considered as "fierce" by the 1977 medium ability students. Item number 26 (Figure 67, chirp) indicates that the 1977 low and medium ability students had a higher probability of success on this item. It is difficult to speculate why this is so: Could it be that Charles Schultz's popular cartoon character, Woodstock, has influenced the response pattern obtained for this item?

For items 33 and 35 (Figures 74 and 76; military and musician, respectively) the 1977 medium ability and the 1956 low ability students had a higher probability of getting this item correct. The differences are probably caused by slightly different discrimination differences between the two study groups as opposed to specific changes in the item difficulty obtained.

Items 34, 36, 41, and 44 (Figures 75, 77, 82, and 85; gypsy, wrestle, medal, and dormitory, respectively) indicate that the 1977 medium ability student had a higher probability of getting the item correct.

Items 37 and 42 (Figures 78 and 83; dwelling and arbor, respectively) indicate that the 1956 high ability students, the 1977 low ability students (in the case of item 37) and the medium ability students (in the case of item 42) had a higher probability of getting these items correct than did the students of corresponding ability level from the other group. Items 25 and 39 (Figures 66 and 80; stumble and slumber) indicate that the 1956 medium ability students had a higher probability of answering these items correctly than did the 1977 medium ability students. Finally, item 43 (Figure 84; garrison) indicates that the 1956 high ability students had a higher probability of answering this item correctly.

Table 30 presents a summary of how the 1977 item parameters differ in relation to the 1956 parameters for each of the aberrant items. A chi square of 0.10 (df=1)

Table 30

The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for Gates Word Recognition

		Discrimination (a)	
		1977 < 1956	1977 > 1977
Difficulty (b)	1977 < 1956	25, 29, 40, 41	20, 26, 34, 36, 44, 45, 48
	1977 > 1956	24, 37, 42	6, 17, 31, 33, 35, 39, 43

suggests that there was no consistent item parameter difference pattern found between the two study groups.

While there did not appear to be any systematic item differences found between the two groups of students, those items that were found to be aberrant did appear to favor the 1977 students when the mean ability differences were calculated. For instance, Table 18 indicated that the mean achievement difference between the 1956 students and the 1977 students was 0.391 ability units. However, the mean achievement difference between the two groups drops when the data from the aberrant items were removed: i.e., when the data from the item that had an estimated item difference that was greater than 0.15 were removed, then the mean achievement between the 1956 and the 1977 students was 0.339 abilty units. This value drops further to 0.296 and then to 0.239 ability units when the data from the items that had an estimated item difference of greater than 0.10 and 0.05, respectively, were removed.

These results suggest that the 1977 students were better able to recognize the words that were administered via the Gates Word Recognition test than did the 1956 students, even after the data from the aberrant items were removed and even though the the 1956 students were older, on the average, than the 1977 students.

California Mental Maturity Test

Items 6, 26, 34, 35, 37, 38, 43, 44, 47, 48, 49, 50, 51, 52, 58, 60, 61, 66, 67, 68, 71, 72, 73, 74, 75, 76, 77, 78, 79, 81, 82, 83, 84, 85, 86, 88, 91, 92, and 98 were identified from Table 20 as having a difficulty (b) parameter that was less than -4.00 or greater than +4.00, as calculated from either the 1956 or the 1977 data. The item characteristic curves for those items that had a difficulty (b) parameter of less than -4.00 would take the form of Figure 1. The item characteristic curves for those items that had a difficulty (b) parameter of greater than +4.00 would take the form of Figure 148. These items were generally so difficult or so easy that students from both study years either got them correct in the case of the easy items, or wrong in the case of the difficult items.

Items 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 27, 28, 29, 30, 31, 32, 33, 36, 40, 42, 45, 55, 56, 59, 62, 64, 80, 87, 93, and 97 (Figures 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 122, 124, 125, 129, 130, 132, 133, 135, 139, 140, 143, and 147, respectively) had estimated item differences of less than 0.05, i.e., these items displayed no evidence of item bias between the two groups of students. In other words, examinees of a given ability level in either 1956 or in 1977 would appear to have an equal probability of answering these items correctly.

The item characteristic curves (Figures 89 to 147), in conjunction with the results presented in Tables 23 and 31 form the basis of the following discussion of where the location of the achievement differences between the 1956 and

1977 students occured when the examinee ability continuum was roughly divided up into three parts: low ($\theta < -1$), medium ($-1 < \theta > +1$), and high ($\theta > +1$).

Items 1, 2, 3, 4, 7, 8, 9, 10, 90, and 96 (Figures 89, 90, 91, 92, 94, 95, 96, 97, 142, and 146, respectively) had an estimated item difference of greater than 0.15. Items 1 through 10 deal with the examinee's ability to identify the left or the right limbs of various given figures. The medium ability 1956 students had a higher probability of getting items 1, 2, 3, and 4 correct than did the 1977 students of similar ability. This is also true of the 1956 high ability students for items 1, 2, 3, 4, 7, 8, 9, and 10. Generally, these results would suggest that the 1956 students were better able to correctly identify the right or left limbs of the given figures than were the 1977 students; except for low and medium ability 1977 students for items 7, 8, 9, and 10. The reason why there was an interaction effect identified between the 1956 and the 1977 students on items 7, 8, 9, and 10 was because these items did not discriminate the good students from the poor students as well in 1977 as they did in 1956. Items 90 and 96 were two vocabulary questions (athletic and construction) that the 1977 students of all ability levels had a higher probability of answering correctly than did the 1956 students.

Items 54, 69, and 89 (Figures 128, 137, and 141, respectively) had an estimated item difference of greater than 0.10, but less than 0.15. Item 54 was a numerical

Table 31

Location of Item Bias by Ability Level for the California Mental Maturity Test

	Ability Level					
	1956			1977		
item #	low	med.	high	low	med.	high
1		*	*			
2		*	*			
3		*	*			
4		*	*			
5		*				
6						
7			*		*	
8			*	*	*	
9			*	*	*	
10			*	*	*	
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						*
23						
24						
25						
26						
27						
28						
29						

Table 31 (con't)

	Ability Level					
	1956			1977		
item #	low	med.	high	low	med.	high
30						
31						
32						
33						
34						
35						
36						
37						
38						
39	*	*				
40						
41	*	*				
42						
43						
44						
45						
46					*	*
47						
48						
49						
50						
51						
52						
53					*	*
54	*	*				*
55						
56						
57					*	*
58						
59						

Table 31 (con't)

Ability Level

	1956			1977		
item #	low	med.	high	low	med.	high
60						
61						
62						
63		*				
64						
65			*			
66						
67						
68						
69		*				
70		*	*			
71						
72						
73						
74						
75						
76						
77						
78						
79						
80						
81						
82						
83						
84						
85						
86						
87						
88						
89					*	*

Table 31 (con't)

Ability Level

item #	1956 low	1956 med.	1956 high	1977 low	1977 med.	1977 high
90				*	*	*
91						
92						
93						
94					*	
95		*	*			
96				*	*	*
97						
98						

* Indicates those areas in the ability continuum where the students from a given study year had a higher probability of getting a given item correct than did the students of a similar ability level from the opposing study year.

series question that did not discriminate as well in 1956 as it did in 1977, consequently the lower ability 1956 students performed better than did the lower ability 1977 students, but the higher ability 1977 students had a higher probability of correctly answering this item than the high ability 1956 students. Item 69 instructed the examinees to mark the alternative that contained the number of cookies that would occur if the total number of cookies were divided among three persons. While the item was not a good discriminator for either study year, the 1956 students had a higher probability of answering this item correctly than did the 1977 students. Item 89 was a vocabulary item (vehicle) on which the 1977 students had a higher probability of getting correct than did the 1956 students.

Items 5, 22, 39, 41, 46, 53, 57, 63, 65, 70, 94, and 95 (Figures 93, 109, 121, 123, 126, 127, 131, 134, 136, 138, 144, and 145, respectively) had an estimated item difference of greater than 0.05, but less than 0.10. Item 5 was one of the right/left identification problems which the 1956 medium ability students had a higher probability of answering correctly. Item 22 was a spatial relation problem that the 1977 high ability students had a higher probability of answering correctly than did the 1956 high ability students. Items 39, 41, and 46 dealt with item relationships. The discrimination parameters (a) for these items indicated that discrimination was not good and that guessing was high for both groups of students. However, the 1956 low and medium ability students, for items 39 and 41, and the high ability 1977 students, for item 46, had a higher probability of getting this item correct than did the students of similar ability levels from the other group of students. Items 53 and 57 were numerical series questions which the medium and high ability 1977 students had a higher probability of answering correctly than did the 1956 students. Items 63, 65, and 70 were computational problems where the medium or high ability 1956 students had a higher probability of getting these items correct than did the 1977 students. Items 94 and 95 were vocabulary questions (eclipse and studious, respectively) that the medium ability 1977 students, for item 94, and the medium and high ability 1956 students, for item 95, had a higher probability of answering correctly than did a student of a similar ability level of the other study year.

Table 32 presents a summary of how the 1977 item parameters differed in relation to the 1956 parameters for each of the aberrant items. A chi square of 0.27 (df=1) suggests that there was no consistent item parameter difference pattern found between the two study groups.

There did not appear to be any systematic pattern of item differences found between the two groups of students. Table 24 indicated that the mean achievement difference between the 1956 students and the 1977 students was 0.553 ability units in favor of the 1977 students when all the test items were included in the computation of the mean

Table 32

The Relationship Between the 1977 Aberrant Item Parameters and the 1956 Aberrant Item Parameters for the California Mental Maturity Test

		Discrimination (a)	
		1977 < 1956	1977 > 1956
Difficulty (b)	1977 < 1956	8, 9, 10, 57, 94, 96	22, 46, 53, 89, 90
	1977 > 1956	1, 2, 3, 4, 7, 63, 65, 95	5, 39, 41, 54, 69, 70

abilities. When the data from the the items that had an estimated item difference of greater than 0.15 were removed, prior to analysis, then the mean achievement difference between the 1956 students and the 1977 students was 0.501 ability units. This value changes to 0.505 and then to 0.472 ability units when the data from the items that had an estimated item difference of greater than 0.10 and 0.05, respectively, were removed.

These results suggest that, on the average, the 1977 students were more intellectually mature, as measured by the California Mental Maturity Test, than the 1956 students by about 0.5 ability units. It is interesting to note that this difference was found following the removal of the aberrant items and in spite of the fact that the 1956 students were, on the average, older than the 1977 students.

C. Classical Test Theory Implications versus Latent Trait Theory Implications

The results from Raven's Progressive Matrices test were employed in the following discussion as an example of the differences that were found between the implications of the results obtained by using classical test theory methodology (Table 33)[8] and latent trait theory methodology (Tables 3 and 5). While the results of these two methodologies are not

---

[8] Difficulty (diff.) is the proportion of examinees who correctly answered a given test item and the discrimination coefficient (disc.) is the point biserial correlation between a given item response and the total test score on all the other items in the test.

Table 33

Classical Test Theory Item Statistics for Raven's Progressive Matrices Test

	1956		1977		delta	
item #	diff.	disc.	diff.	disc.	diff.	disc.
1	0.999	0.008	0.995	0.131	0.004	-0.123
2	0.999	0.019	0.995	0.115	0.004	-0.096
3	0.998	0.036	0.984	0.191	0.014	-0.155
4	0.994	0.040	0.980	0.243	0.014	-0.203
5	0.986	0.077	0.972	0.244	0.014	-0.167
6	0.983	0.081	0.971	0.240	0.012	-0.159
7	0.757	0.331	0.842	0.364	-0.085	-0.033
8	0.803	0.215	0.770	0.236	0.033	-0.021
9	0.850	0.310	0.857	0.372	-0.007	-0.062
10	0.806	0.306	0.815	0.378	-0.009	-0.072
11	0.447	0.323	0.437	0.332	0.010	-0.009
12	0.262	0.296	0.309	0.291	-0.047	0.005
13	0.993	0.090	0.978	0.317	0.015	-0.227
14	0.985	0.101	0.963	0.304	0.022	-0.203
15	0.981	0.052	0.956	0.295	0.025	-0.243
16	0.725	0.423	0.846	0.398	-0.121	0.205
17	0.713	0.395	0.787	0.430	-0.074	-0.035
18	0.752	0.481	0.770	0.485	-0.018	-0.004
19	0.744	0.431	0.824	0.410	-0.080	0.021
20	0.667	0.437	0.677	0.446	0.000	-0.009
21	0.463	0.426	0.646	0.429	-0.183	-0.003
22	0.556	0.363	0.669	0.376	-0.113	-0.013
23	0.473	0.304	0.614	0.355	-0.141	-0.051
24	0.198	0.335	0.372	0.306	-0.174	0.029
25	0.989	0.065	0.974	0.249	0.015	-0.184
26	0.970	0.174	0.956	0.316	0.014	-0.142
27	0.894	0.367	0.918	0.383	-0.024	-0.016
28	0.826	0.437	0.892	0.399	-0.066	0.038
29	0.625	0.476	0.732	0.445	-0.107	0.031
30	0.473	0.446	0.683	0.394	-0.210	0.052
31	0.429	0.397	0.508	0.311	-0.079	0.086
32	0.293	0.460	0.514	0.456	-0.221	0.004
33	0.347	0.487	0.569	0.466	-0.222	0.021
34	0.414	0.477	0.688	0.505	-0.274	-0.028
35	0.288	0.482	0.556	0.501	-0.268	-0.019
36	0.149	0.351	0.267	0.345	-0.118	0.006

directly comparable, because they are reported in different metrics, the implications of these results are comparable.

If a difference level of 0.05 between the item difficulties obtained from the 1956 data and the 1977 data is assumed to be representing a significant difference between the two groups of examinees, then the item difficulties calculated, via classical test theory methodology, from the data obtained from items 7, 12, 16, 17, 19, 21, 22, 23, 24, 28, 29, 30, 31, 32, 33, 34, 35, and 36 (see Table 33) would be considered to be different for each examinee group (1956 or 1977). In comparison, items 8, 11, 12, 16, 18, 20, 21, 23, 24, 30, 31, 32, 33, 34, 35, and 36 were identified as being different, between the two study years, by the latent trait methodology employed in this study (see Table 5).

There are four categories by which the classical test theory results and the latent trait results can be compared:

(a) both methods identified a given item as behaving in a similar manner each time the item was administered to each group of examinees;

(b) both methods identified a given item as behaving differently each time the item was administered;

(c) latent trait theory methodology identified a given item as behaving differently each time that the test was administered, but classical test theory methodology identified the item as behaving in a similar manner each time the item was administered; and

(d) classical test theory methodology identified a given item as behaving differently each time that the test was administered, but latent trait theory methodology identified the item as behaving in a similar manner each time the item was administered.

While category (a) is perhaps the least interesting, it is important to note that the only time that there was complete agreement between the two methods was when the items were found by both groups of examinees to be very easy. This does, however, make sense because nearly all of the examinees at all ability levels got these items correct. Consequently, there was little discrimination among the examinees for either analysis procedure to detect significant differences. It would also be reasonable to suspect agreement between the two methods if an item were extremely difficult and guessing were not influencing the data. If guessing were influencing the data, the latent trait theory methodology would be more sensitive to differences between the two groups (if the three parameter latent trait model were employed) than the classical test theory methodology would be and consequently, there would not be agreement between the two methodologies.

Category (b) occurred with items 12, 16, 21, 23, 24, 30, 31, 32, 33, 34, 35, and 36. While the two methods of analysis did agree on the fact that these items behaved differently each time that they were administered, the conditions which led to the identification of these items were different: i.e., the item difficulty index that was obtained by the classical test theory methodology represents the average difficulty of examinees across all levels of examinee ability, whereas the difference value obtained from the latent trait methodology employed in this study represents a weighting of:

1. the probability of success of the examinees of each ability level;
2. the frequency differences found among the various ability levels; and
3. the sensitivity to differences in discrimination values for a given item between the two study groups.

In addition, it is possible to evaluate the probability of success for each examinee ability level through the item characteristic curves. So, the underlying approach of the two methods is quite different, even though they came to the same conclusions for these items.

Category (c) exists for items 8, 11, 18, and 20 (Figures 3, 6, 10, and 12, respectively). The pattern of differences is similar for each of these items. The classical test theory methodology is insensitive to the fact that a greater number of examinees in the medium ability range had a greater difference of probability of success between the two study groups (1956 and 1977) than did the examinees who had either high or low abilities. Consequently, classical test theory methodology would not identify these items as behaving differently on each of the two different test administrations, while latent trait theory methodology would.

Items 7, 17, 19, 22, 28, and 29 (Figures 2, 9, 11, 14, 18, and 19, respectively) can be located in category (d). Once again, the pattern of differences of item difficulty between the two study years as they are measured by the two different methodologies is similar for these items: i.e., the classical test theory methodology is not sensitive to the fact that there are fewer examinees found at the extreme ability levels. Consequently, classical test theory methodology is more likely to be influenced by differences in the extreme ability levels than is latent trait theory methodology.

The implications of the classical test theory methodology and the latent trait theory methodology, as they were applied to these data, would have likely been more similar (except that latent trait theory methodology also allows the researcher to assess the probability of success for each examinee ability level) if the Rudner (1977) method of determining item bias were employed: i.e, the difference between the item characteristic curves that was obtained by weighting the probablity of success of a given examinee ability level by the frequency of that ability, may have been a more sensitive indicator of item bias across different administrations of the same test, than would have been the other methods of analysis that are currently available.

D. Conclusions

Latent trait theory successfully provides for a methodology which can assess the validity of comparing scores that were obtained from a test-retest design; whether or not the retest was administered to the same population or sample as was the original test,[9] in order to assess achievement differences.

The methodology advocated by this thesis, in order to determine which items should be included in the calculation of a mean difference score, is as follows:

1. Calculate the item parameters for each item for each test administration.
2. Determine if any of the items have difficulty parameters that are greater than +4.00 or less than -4.00. If there are, then these items should not be included in the parameter rescaling procedure that follows.
3. Rescale the item parameters from the two test administrations so that they are on the same scale.
4. Calculate the probable item difference score by multiplying the differences, between the two test results, of the probability of success for a given ability level by the proportion of total scores found at that ability level, then summing across all ability

---

[9] Although not demonstrated in this thesis, latent trait theory does not necessarily presuppose that the two tests were identical, but rather that they measure the same latent trait (Lord, 1977).

levels (see equation 7).

5. Determine a level of probable item differences that would constitute an aberrant item.
6. Remove the data obtained from the aberrant item(s).
7. Calculate the examinee abilities (from the combined test-retest data) using the revised data.
8. Calculate the mean group ability from each test administration.
9. Determine the achievement difference between the two test administrations in ability units.

This procedure provides the researcher with several benefits because it allows the researcher to be able to:

1. visually assess how the difficulty (b) parameter and the discrimination (a) parameter affect the different students across all levels of ability through the production of the item characteristic curves;
2. determine whether or not an item is measuring the same underlying trait each time that it is administered;
3. take into account the number of examinees at each ability level in determining whether or not an item is favoring a given test administration group.

While there are several benefits obtained when using a latent trait anaysis, there are some drawbacks that must also be considered, e.g.,

1. The number of examinees that need to be included in the testing procedure needs to be fairly large in order to assure that there is a representative sample of

examinees located at each ability level.

2. The algorithm used to compute the student abilities and the item parameters requires the use of a large main-frame computer or an array processor. Generally speaking, the casual user has limited access to such machines.
3. At the present time, the robustness of the latent trait models has not been fully investigated, hence it is not known what happens when the underlying assumptions are violated.

In addition, there are two other points of interest that should be considered:

1. The breaking away from classical test theory methodology should be approached cautiously, in that, while the latent trait theory has been shown to be sound, the methods by which it has been implemented are often untested in relation to the underlying assumptions of the original theory (see Lumsden, 1976).
2. The idea of reporting achievement scores in ability units may be confusing to persons outside of the test measurement area. This point may be considered minor by some practitioners, but when it is important to the average teacher; i.e., how many teachers understand what a z-score is, let alone understand the meaning of an ability unit?

At the present time, latent trait theory is being applied to large scale testing projects by researchers with backgrounds in measurement and computer applications. While there is no immediate application of latent trait methodology for the average teacher, the benefits of latent trait methodology can readily be seen in standardized testing programs (e.g. E.T.S.; Lord, 1977) and in item banking programs that are involved in the constitution of tailored tests (e.g. Urry, 1977). As latent trait methodology is further tested and its strengths and weaknesses become known, then latent trait methodology will no doubt continue to become more widely implemented.

With respect to the Edmonton Grade III Achievement Study data, the latent trait methodology proposed in this thesis suggests the following results:

1. In all of the four tests evaluated, there were items that were found to be favoring either the 1956 students or the 1977 students at given examinee ability levels. This would imply that these tests were, in fact, measuring different latent traits when they were administered to each study group. This makes intuitive sense since it would be naive to believe that there would have been no curriculum changes occurring over the last 21 years in Grade III. While latent trait methodology can identify those items that did not behave in the same manner each time that the test was administered, latent trait theory can not tell the researcher why the item behaved differently. The researcher must speculate the cause by investigating

the changes that have occurred with respect to the material presented to the examinee by the item in question: both internally, within respective school curricula, and externally, without the formal school systems.

2. In all of the four tests evaluated, there was no consistent pattern of bias found among the identified aberrant items. Essentially, this implies that, on the average, neither study group was favored more than the other when the tests were administered. However, there were differences with regards to specific items. When the data from these items were removed, prior to analysis, then the following results were found:
    a. the 1977 students obtained a score of 0.46 ability units above the 1956 students on the Raven Progressive Matrices test;
    b. the 1977 students obtained a score of 0.189 ability units above the 1956 students on the Gates Paragraph Reading Test;
    c. the 1977 students obtained a score that was 0.239 ability units above the 1956 students on the Gates Word Recognition test; and
    d. the 1977 students obtained a score that was 0.472 ability units above the 1956 students on the California Mental Maturity Test.

Since the 1956 students were reported by Clarke et al. (1978) as being 2.53 months older, on the average, than the 1977 students, then the differences found between the 1956 and 1977 students would likely increase if the tests had been conducted for the same age groups. These results suggest that with regard to these four tests, the 1977 students, on the average, performed better than the 1956 students. As these groups could be construed as being populations, any difference between them could be considered significant, but are these differences meaningful? The answer rests with what the reader would wish to use these results for.

It is interesting to note that the two "intelligence" tests (the Raven Progressive Matrices and the California Mental Maturity tests) were in close agreement as to what were the mental ability differences between the two study groups (0.46 and 0.472 ability units, respectively). If the premise is accepted that both tests measure "intelligence", then the closeness of the ability differences scores would tend to support the appropriateness of the latent trait methodology employed within this study.

In summary, the methodology described in this thesis appears to be a viable method for determining item bias when different groups of examinees have been administered a given test at different times and this, in turn, allows the researcher to accurately assess achievement differences between the two study groups.

Figures

Figure 1

ICC'S FOR ITEM: b ≤ 4.00

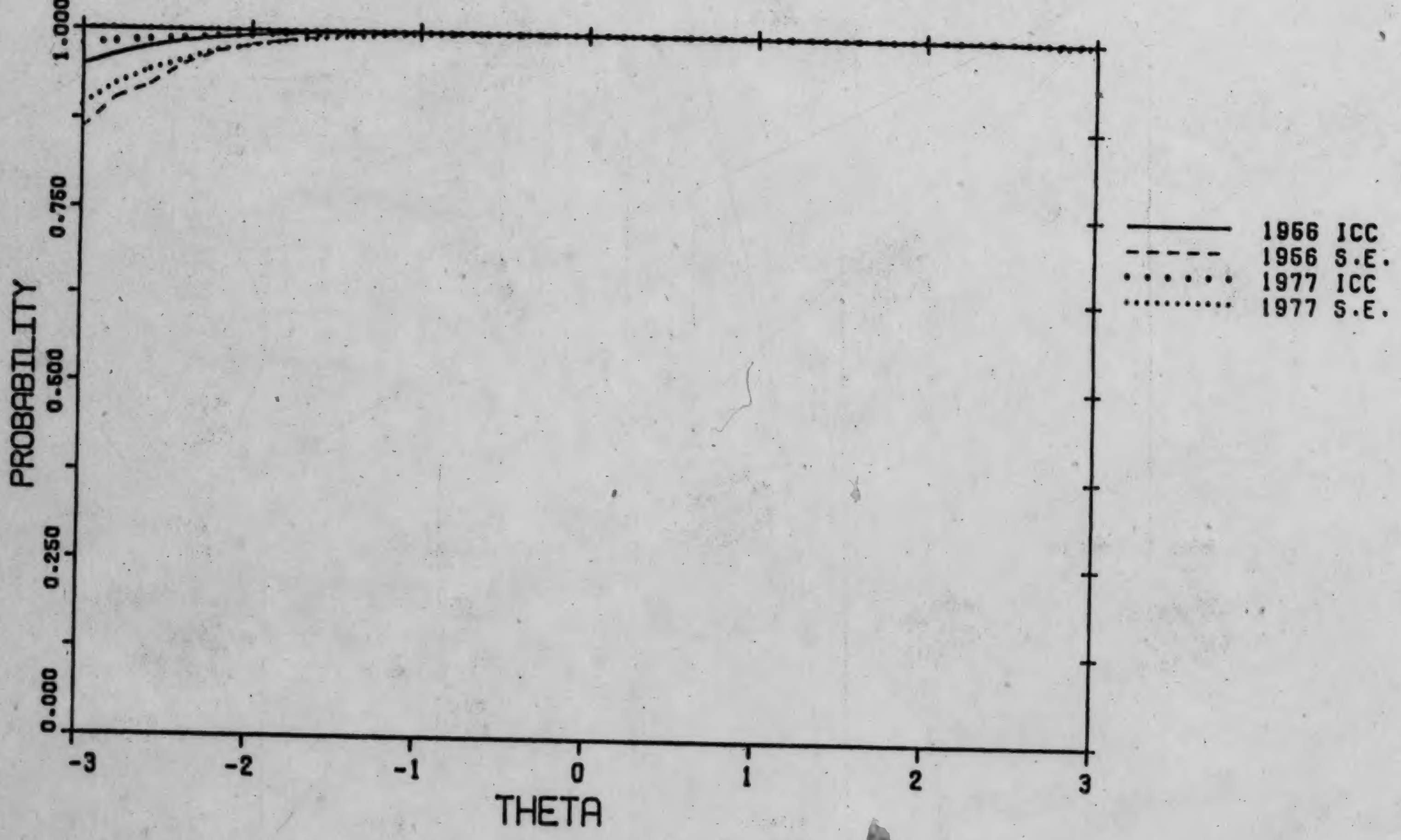


Figure 2

ICC'S FOR ITEM NUMBER 7 : RAVEN

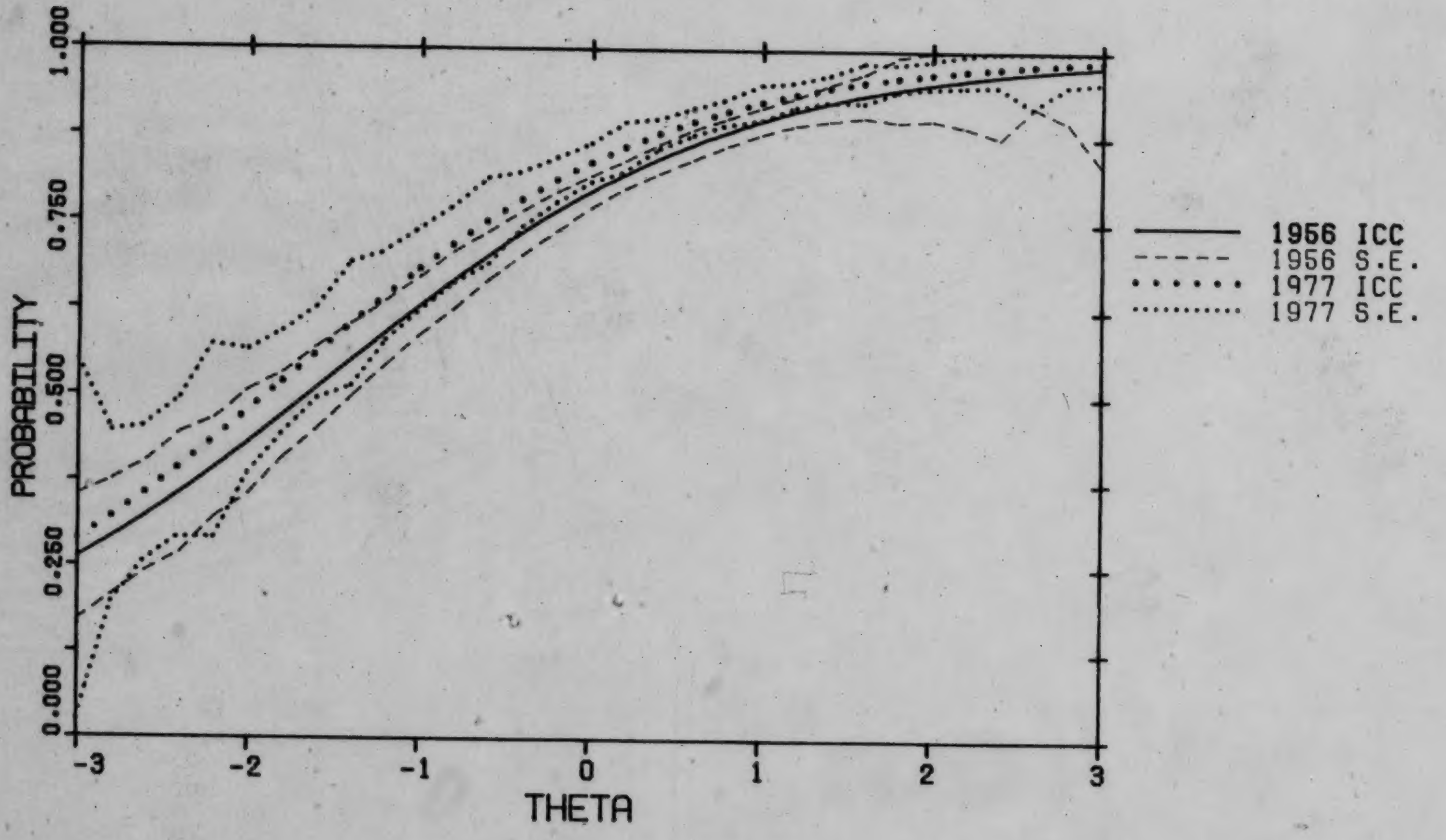


Figure 3

ICC'S FOR ITEM NUMBER 8 : RAVEN

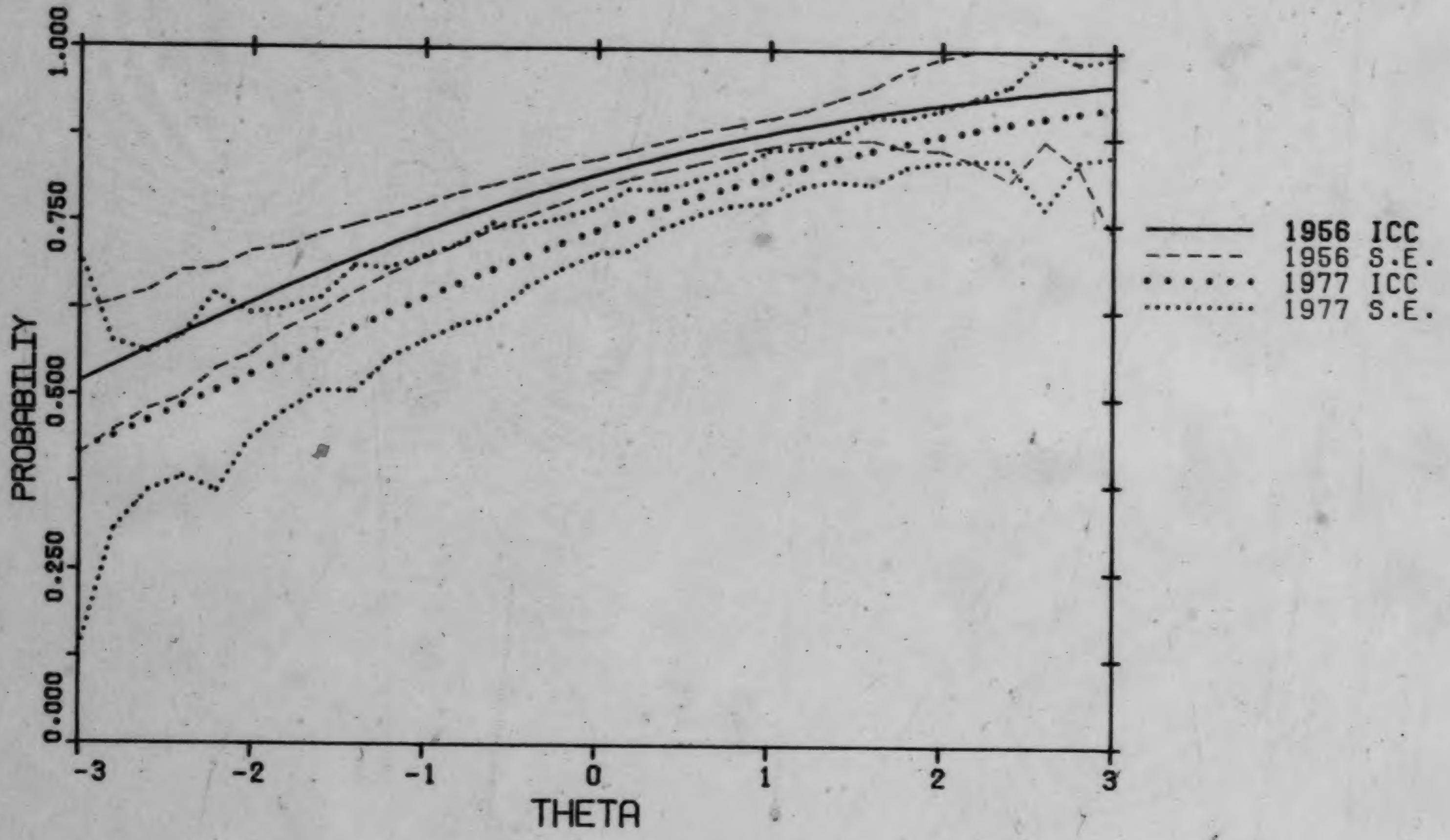


Figure 4

ICC'S FOR ITEM NUMBER 9 : RAVEN

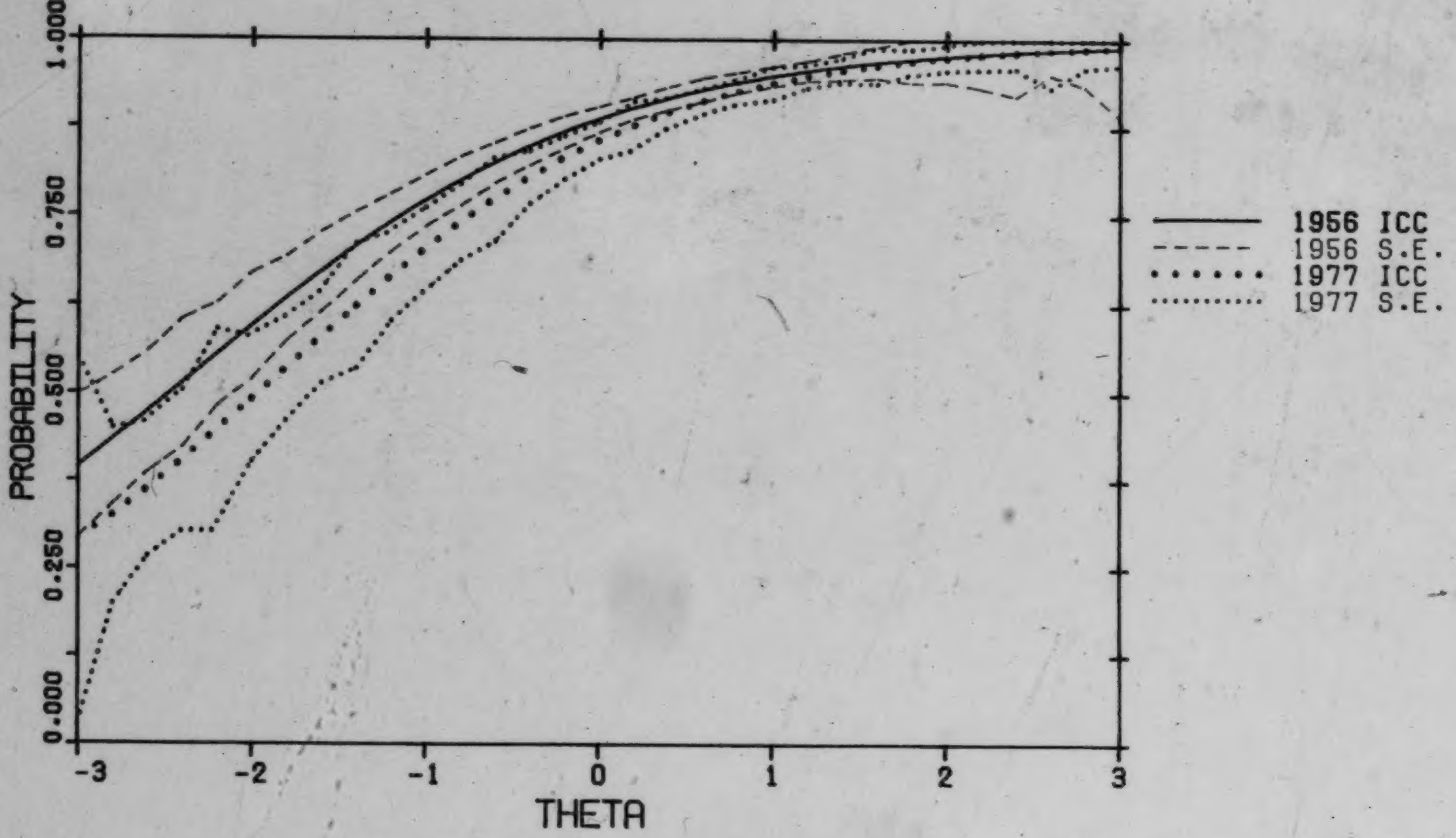


Figure 5

ICC'S FOR ITEM NUMBER 10 : RAVEN

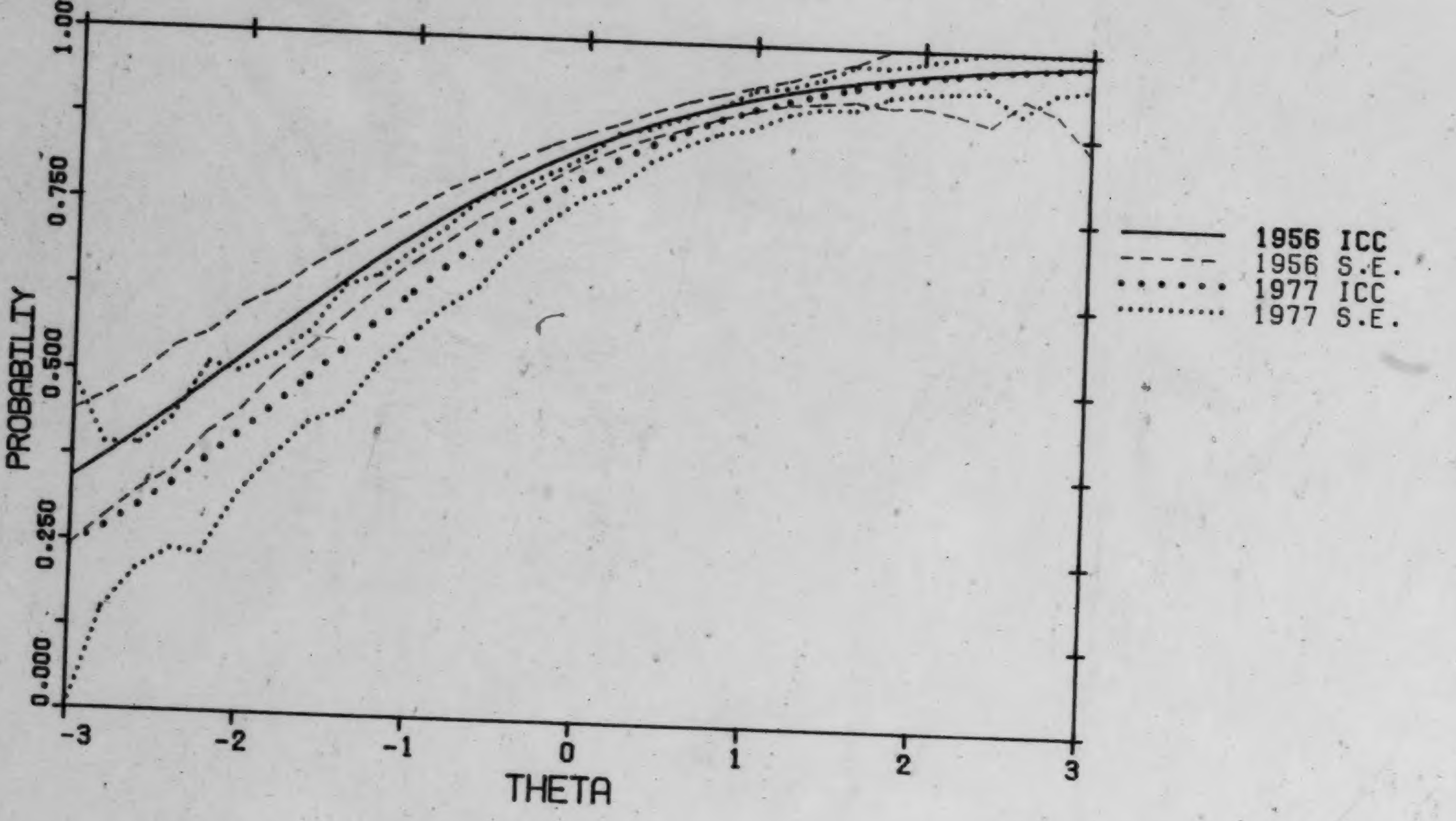


Figure 6

ICC'S FOR ITEM NUMBER 11 : RAVEN

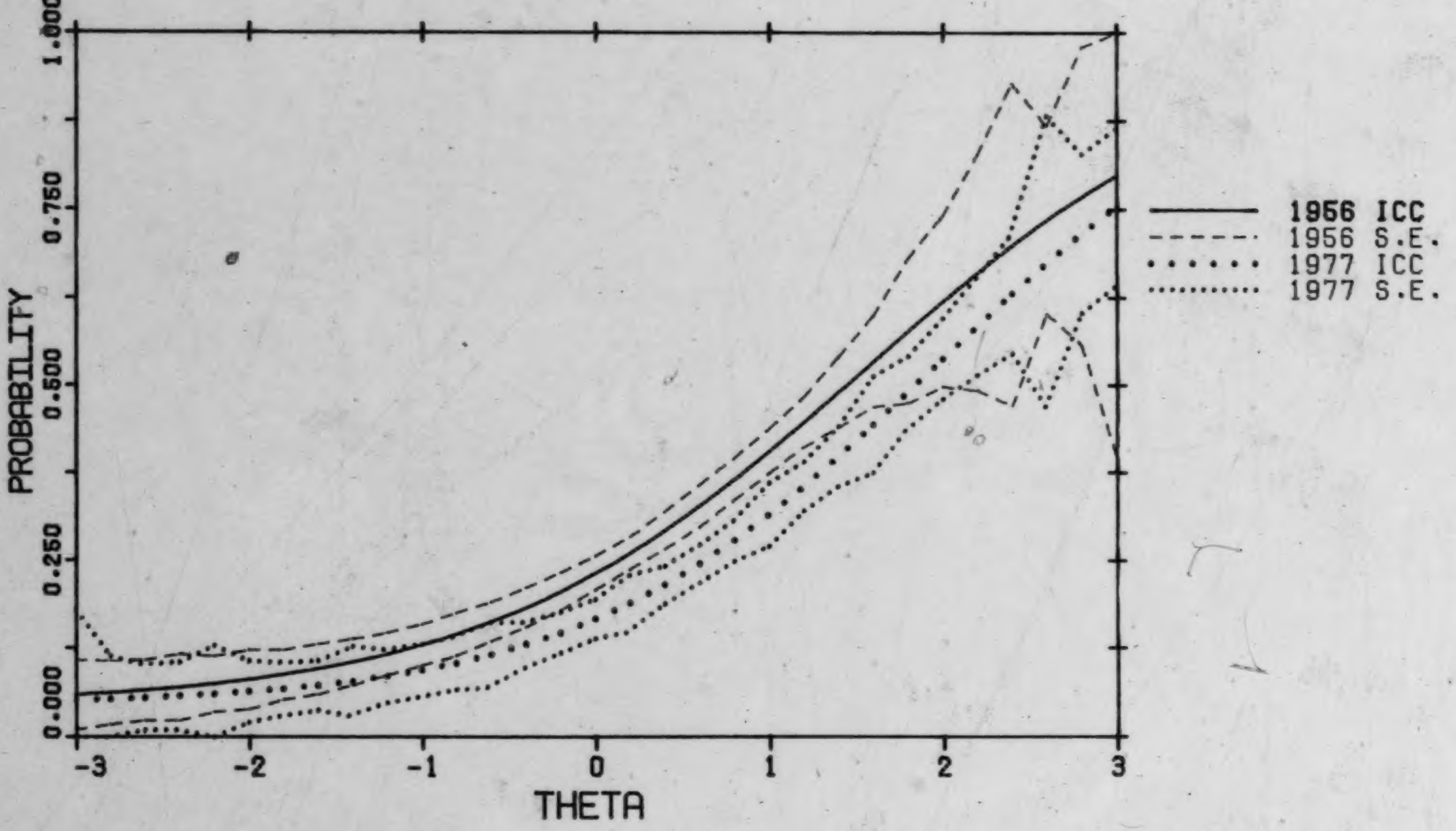


Figure 7

ICC'S FOR ITEM NUMBER 12 : RAVEN

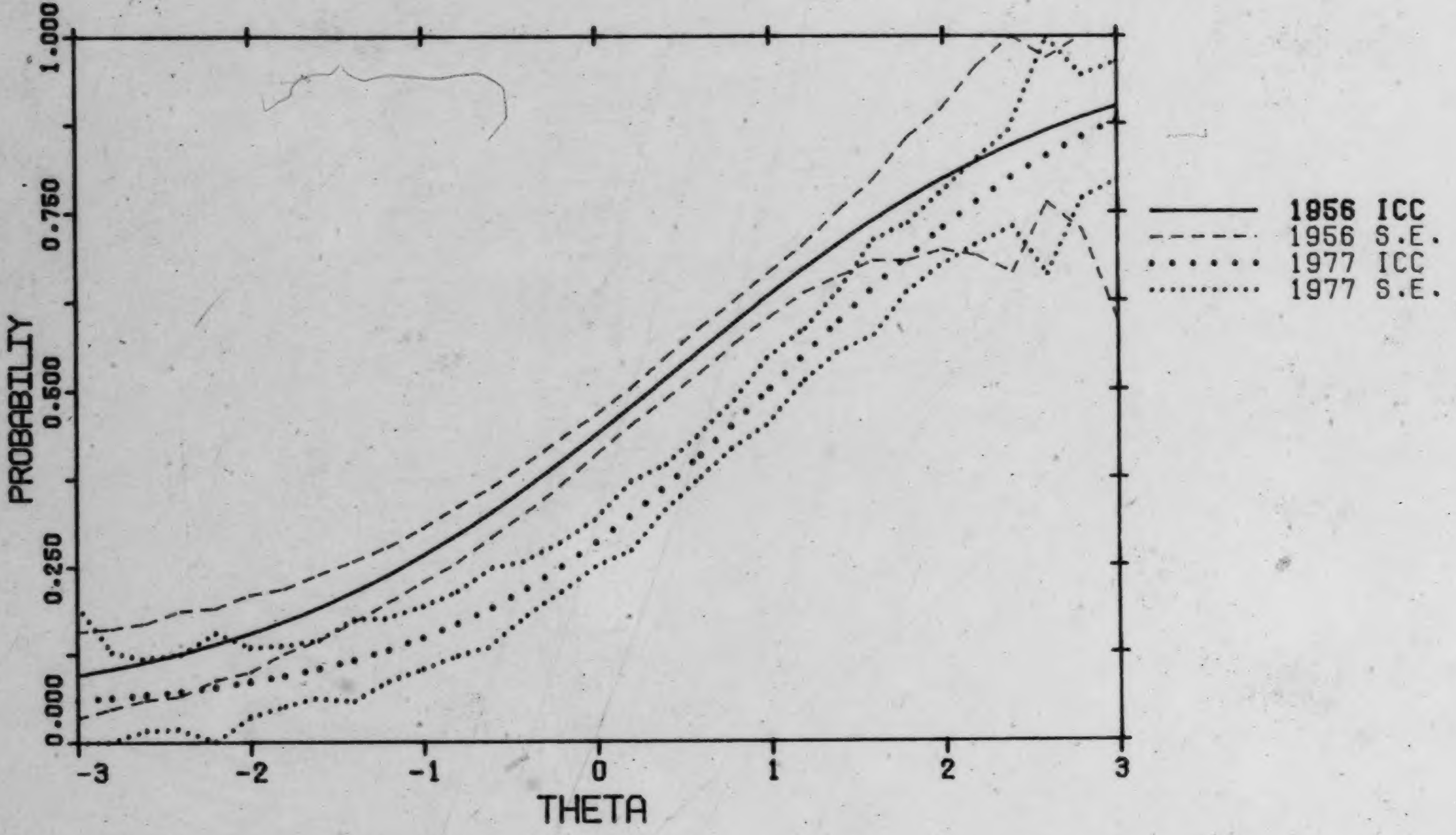


Figure 8

ICC'S FOR ITEM NUMBER 16 : RAVEN

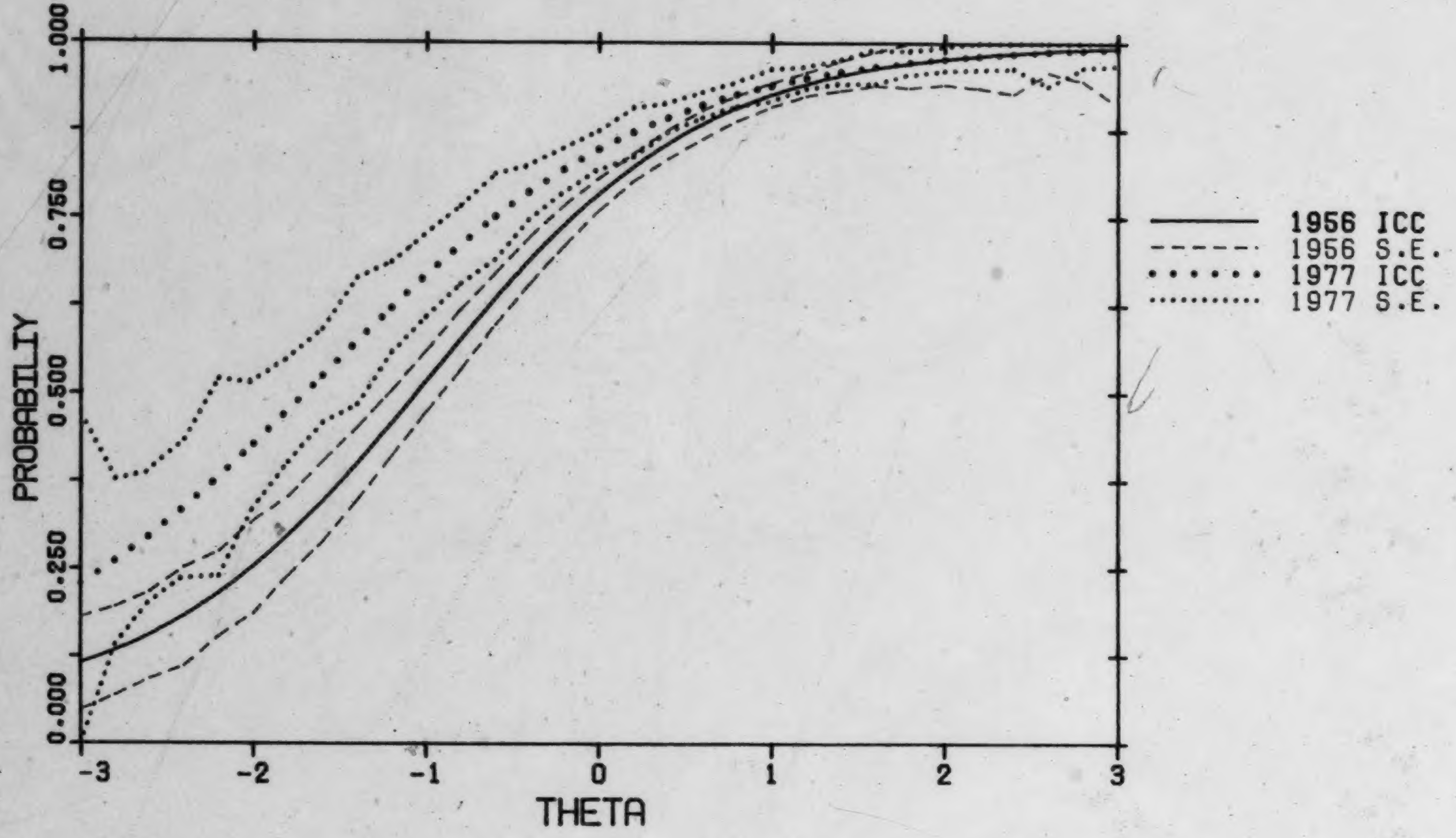


Figure 9

ICC'S FOR ITEM NUMBER 17 : RAVEN

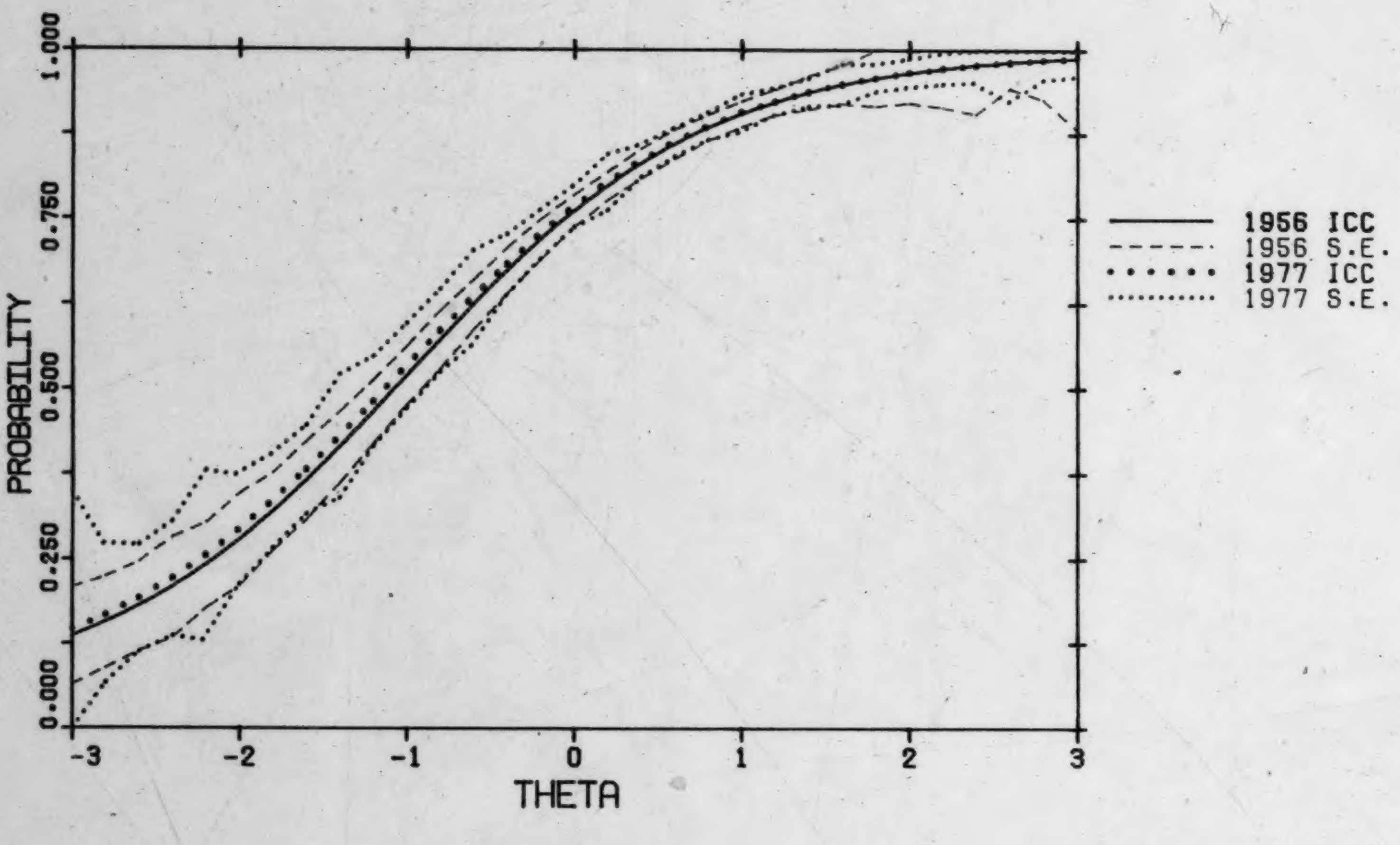


Figure 10

ICC'S FOR ITEM NUMBER 18 : RAVEN

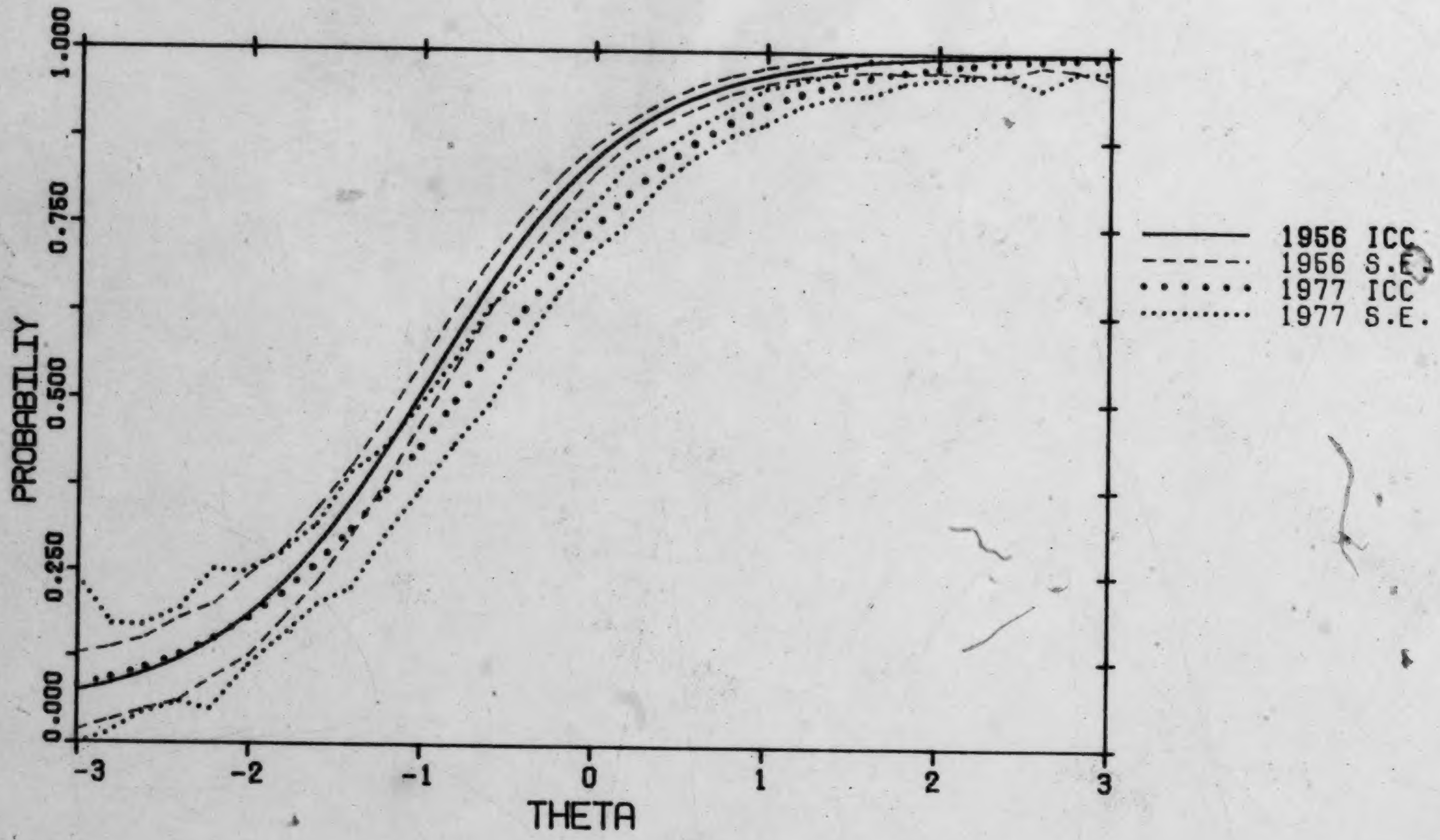


Figure 11

ICC'S FOR ITEM NUMBER 19 : RAVEN

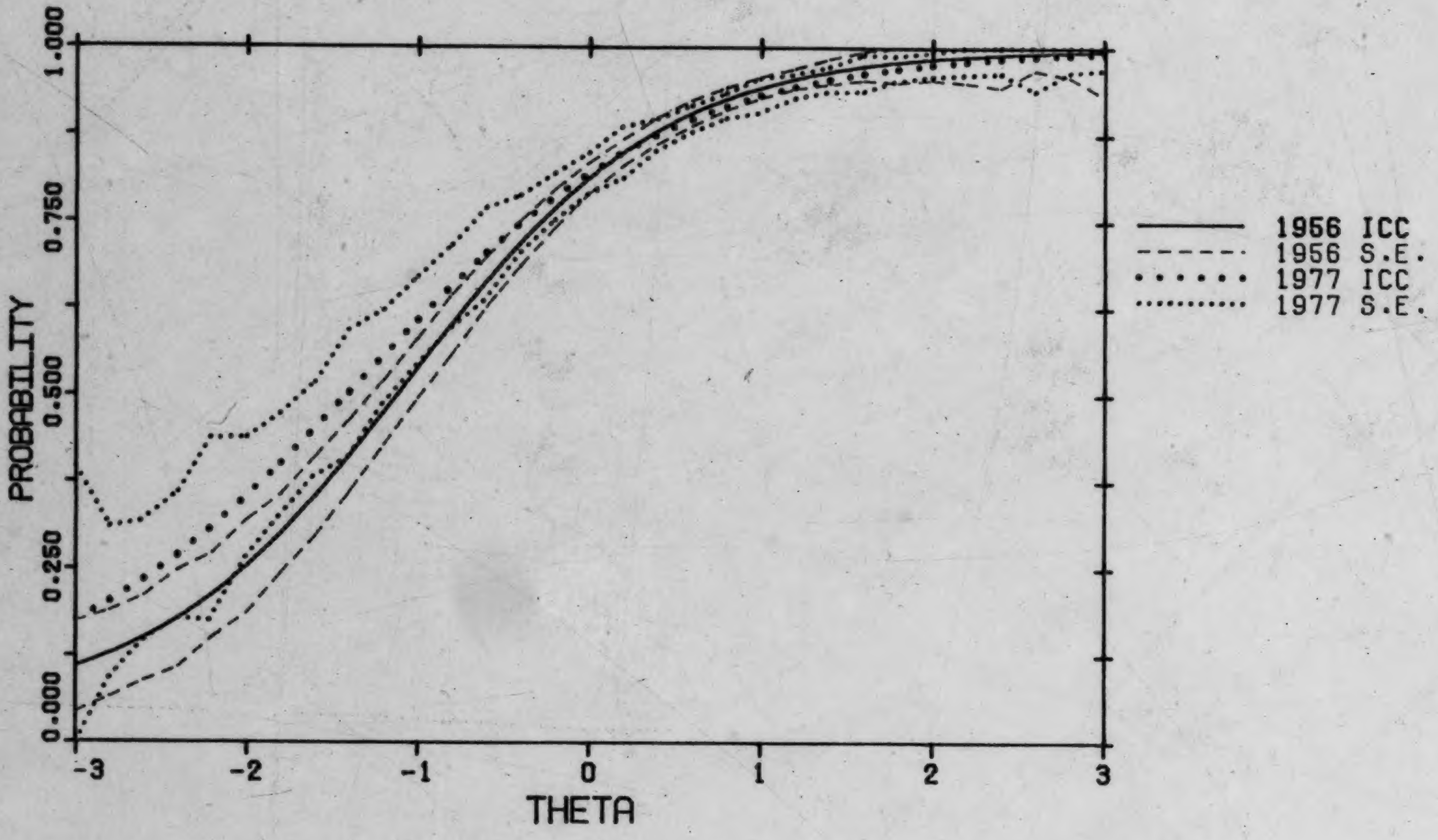


Figure 12

ICC'S FOR ITEM NUMBER 20 : RAVEN

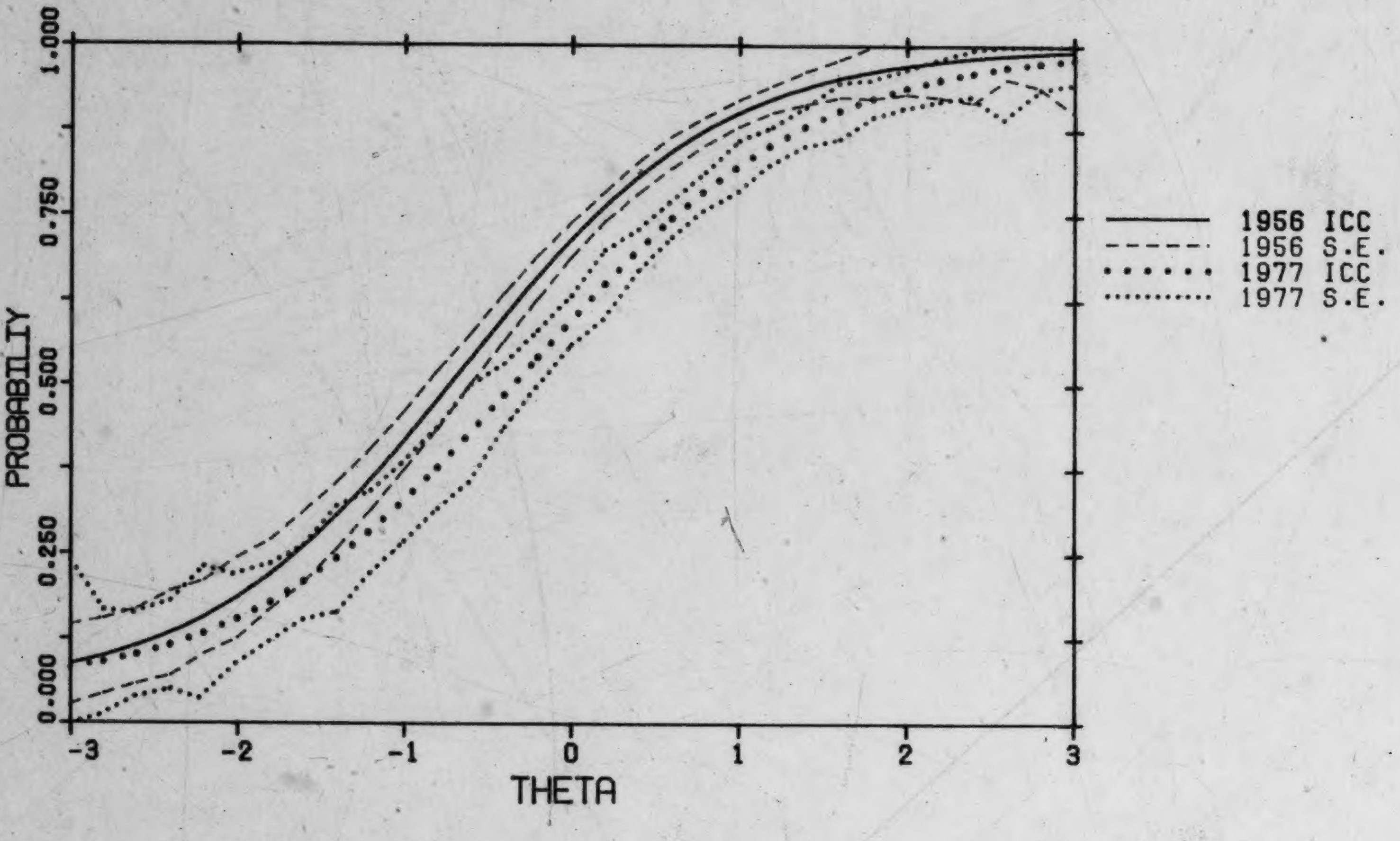


Figure 13

ICC'S FOR ITEM NUMBER 21 : RAVEN

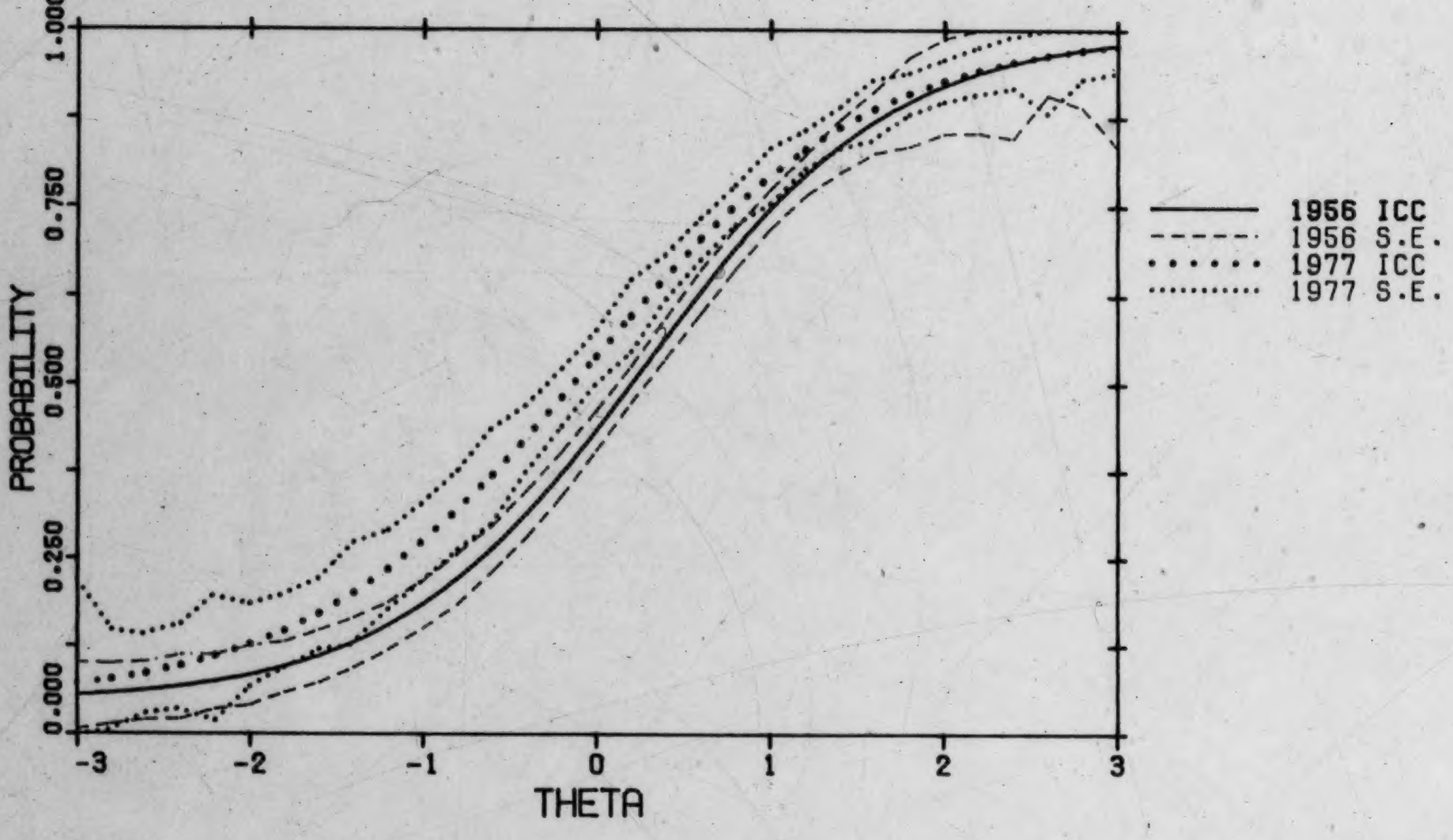


Figure 14

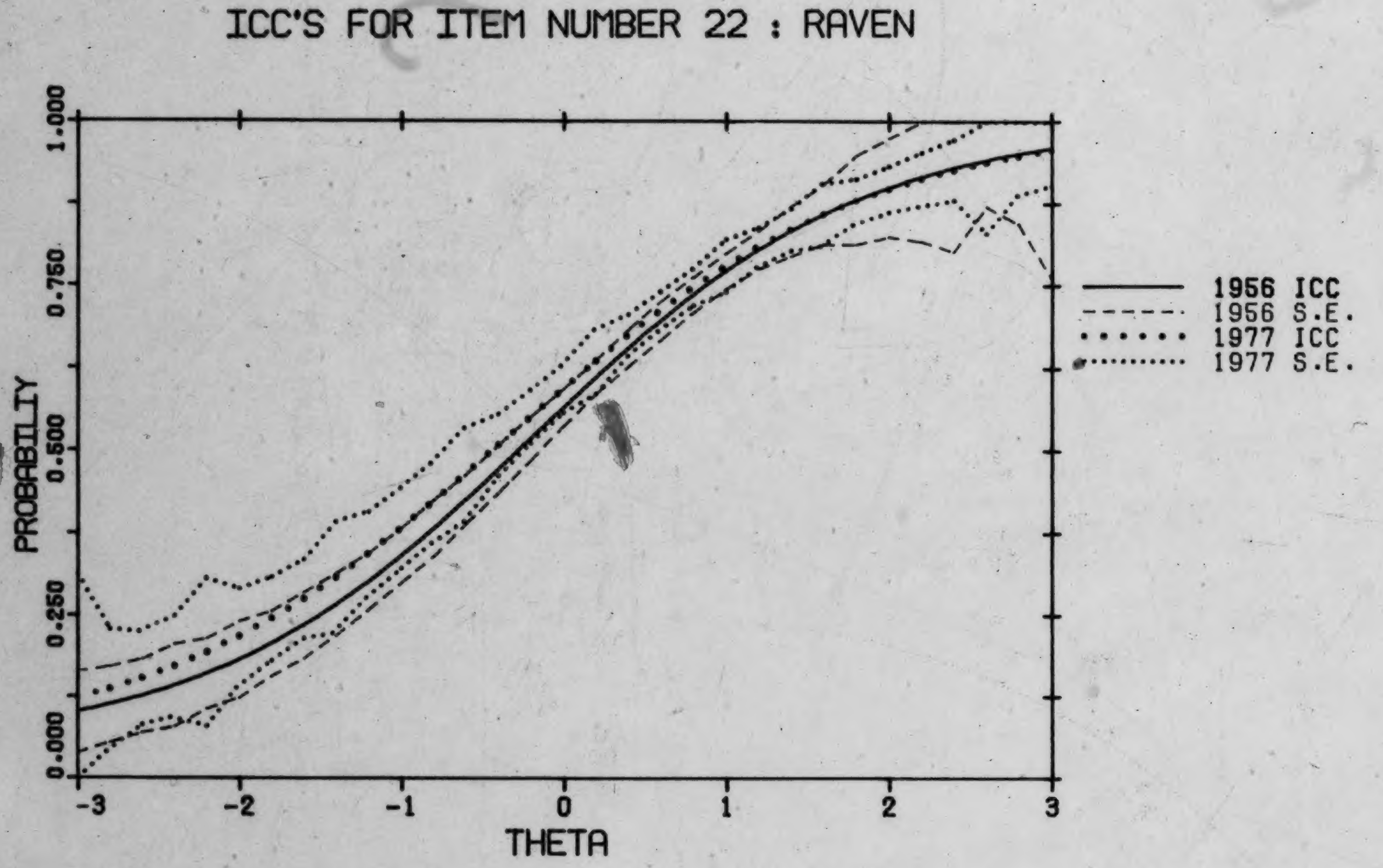

ICC'S FOR ITEM NUMBER 22 : RAVEN
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 15

ICC'S FOR ITEM NUMBER 23 : RAVEN

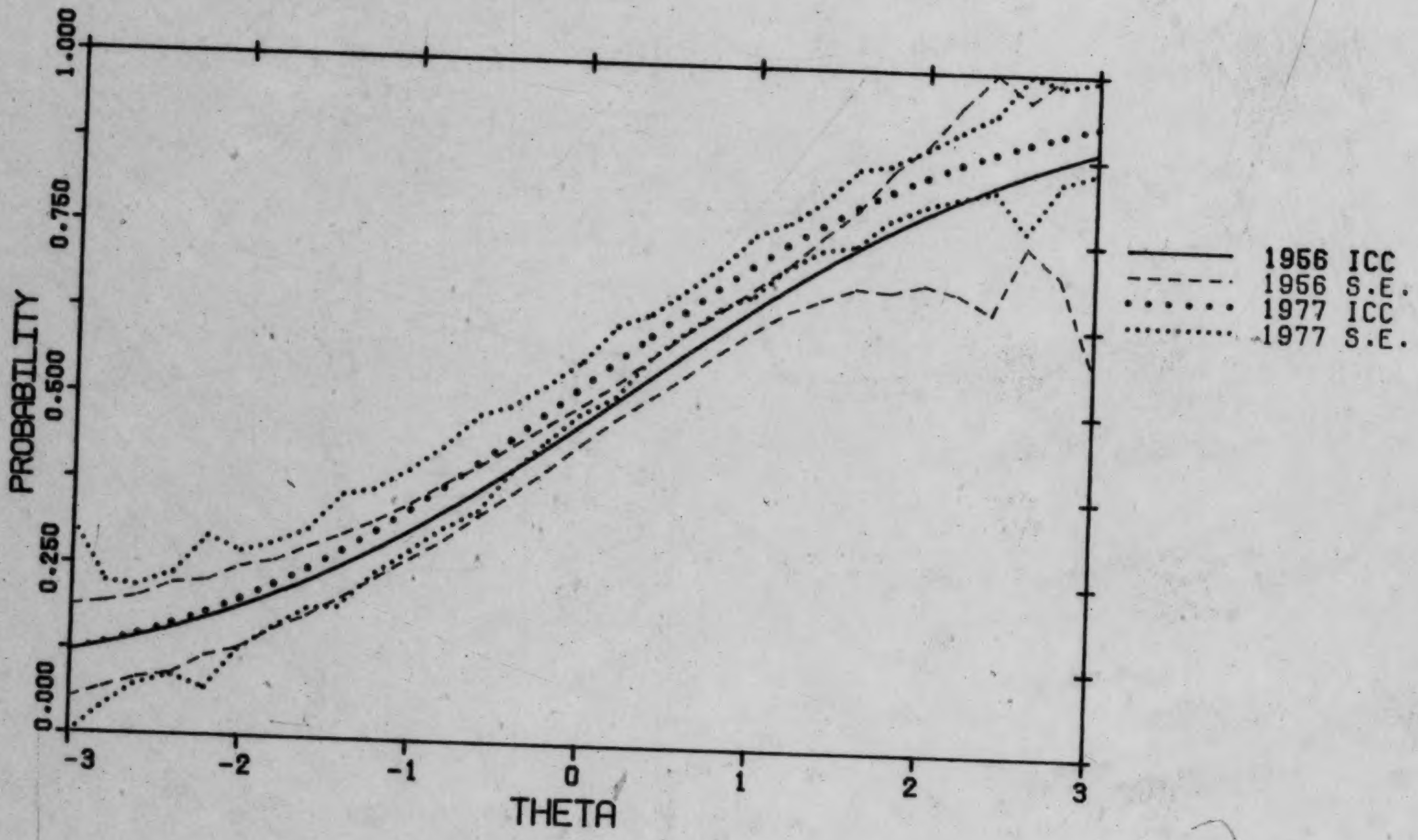


Figure 16

ICC'S FOR ITEM NUMBER 24 : RAVEN

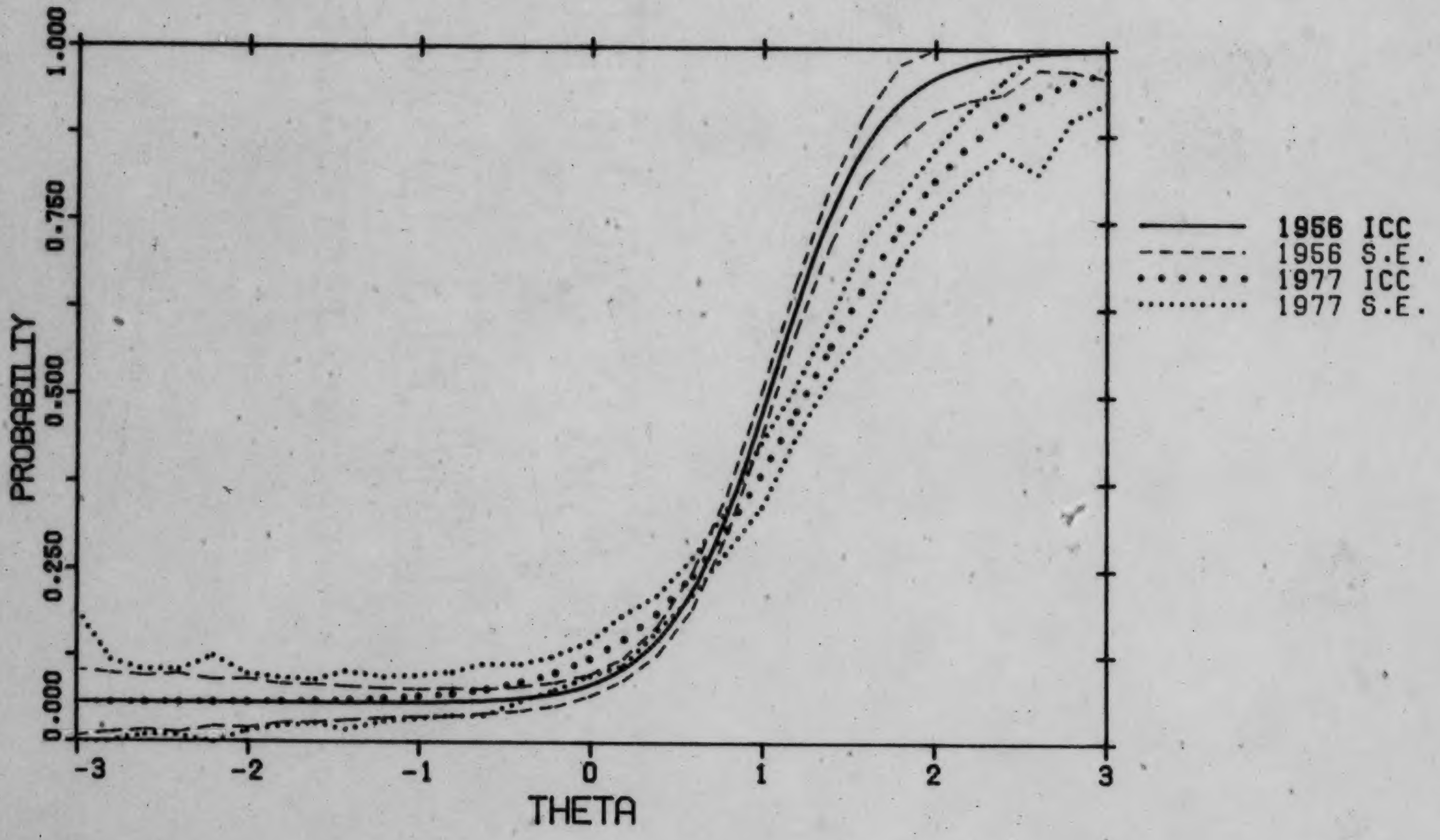


Figure 17

ICC'S FOR ITEM NUMBER 27 : RAVEN

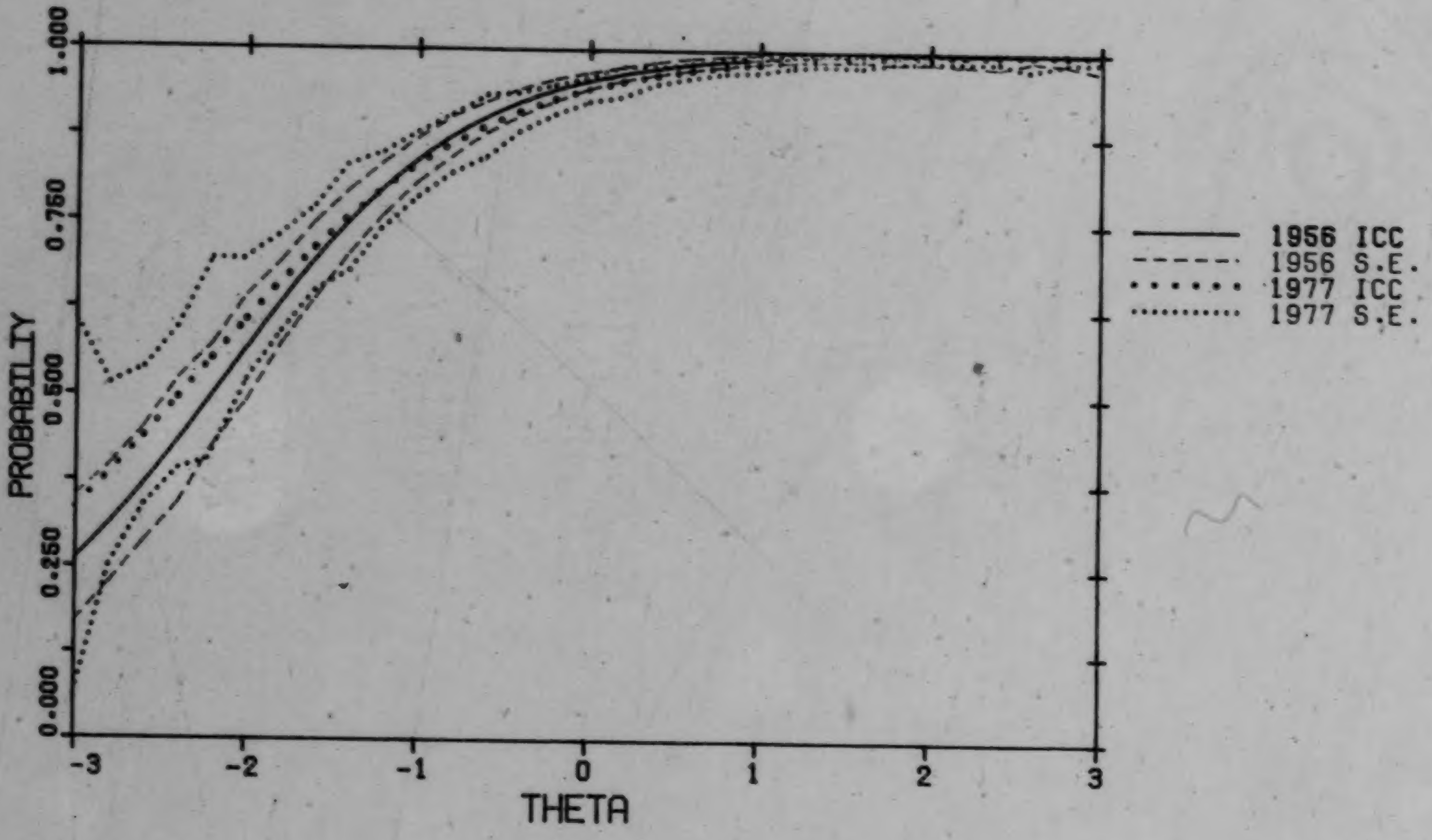


Figure 18

ICC'S FOR ITEM NUMBER 28 : RAVEN

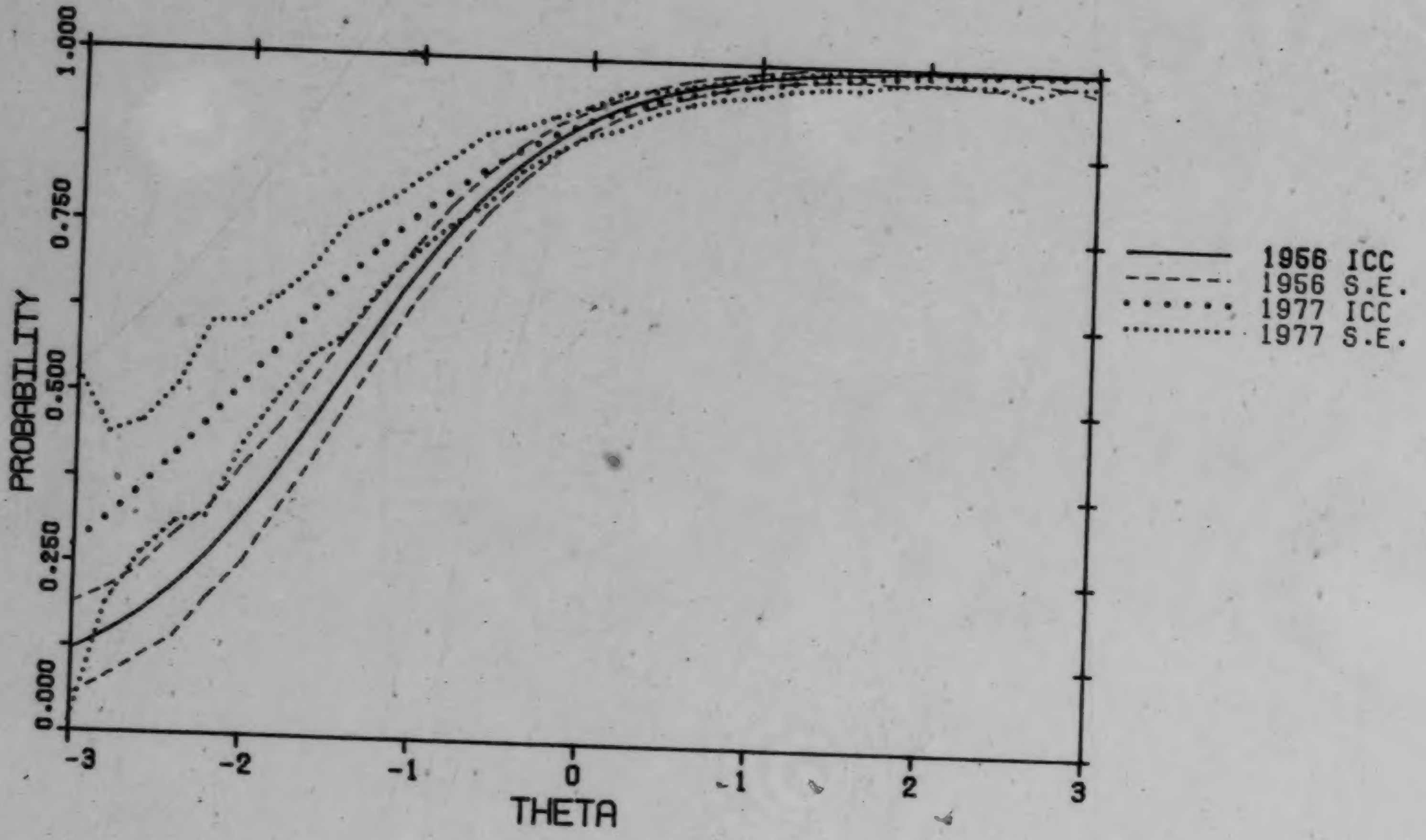


Figure 19

ICC'S FOR ITEM NUMBER 29 : RAVEN

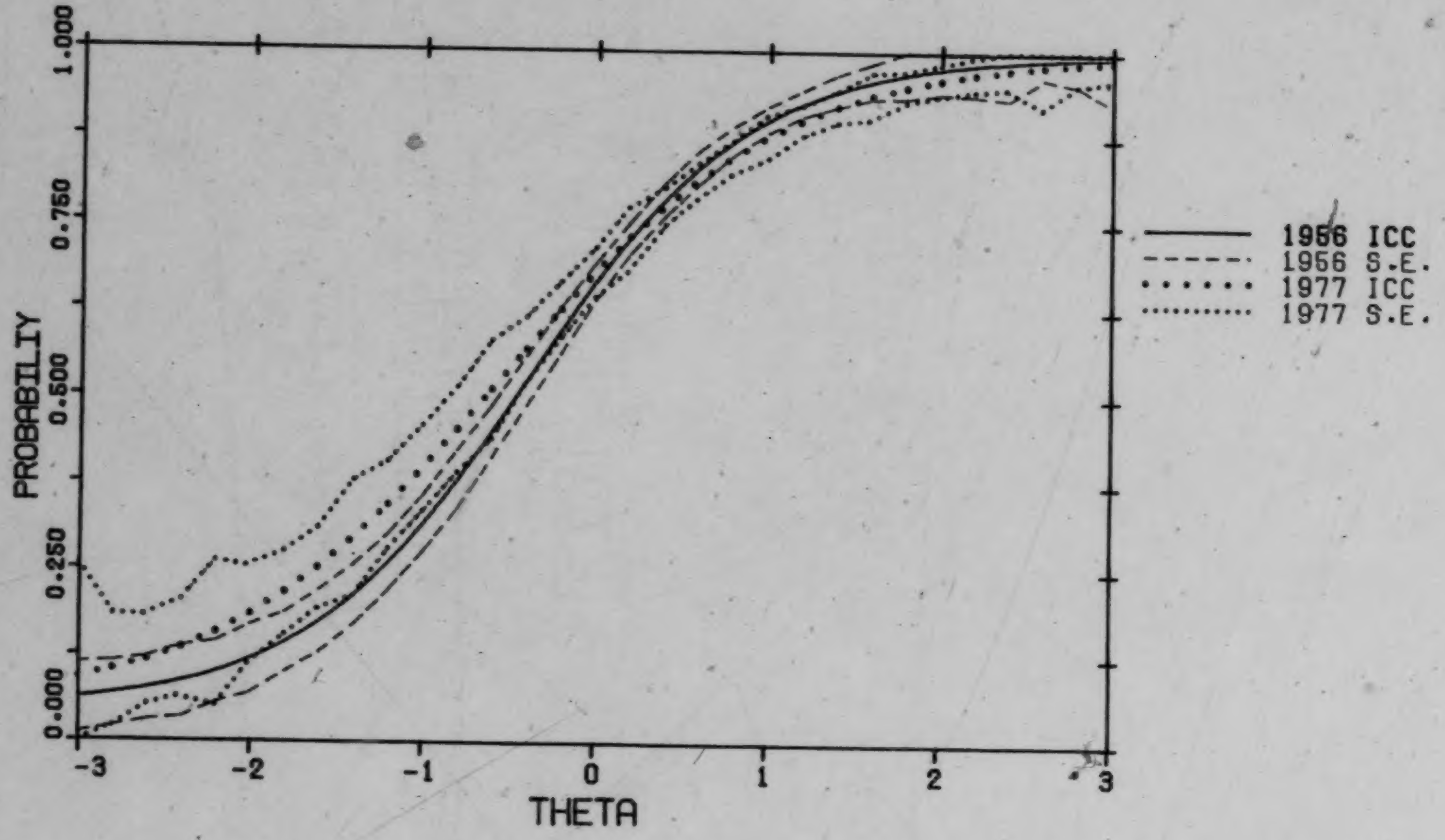


Figure 20

ICC'S FOR ITEM NUMBER 30 : RAVEN

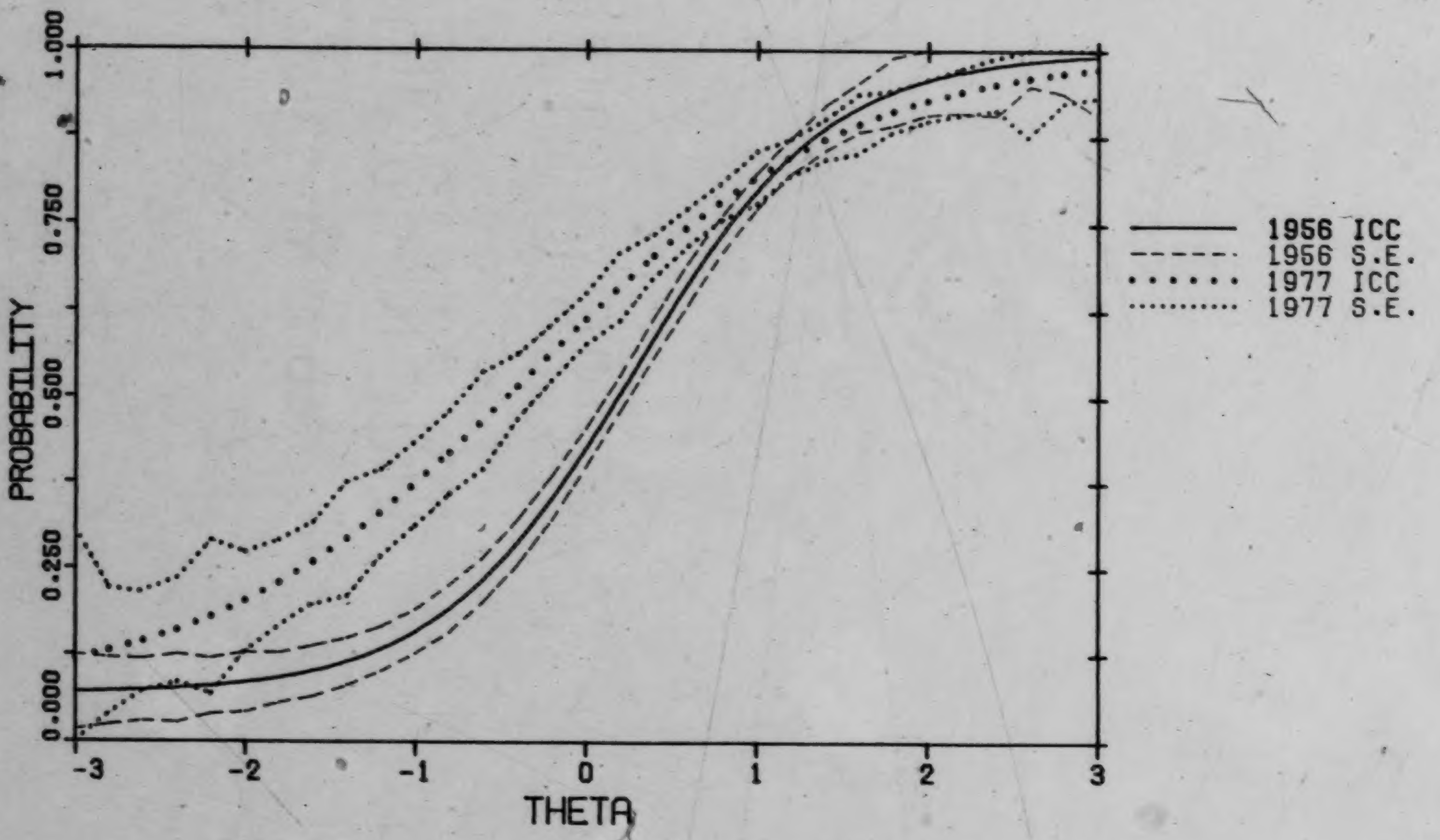


Figure 21

ICC'S FOR ITEM NUMBER 31 : RAVEN

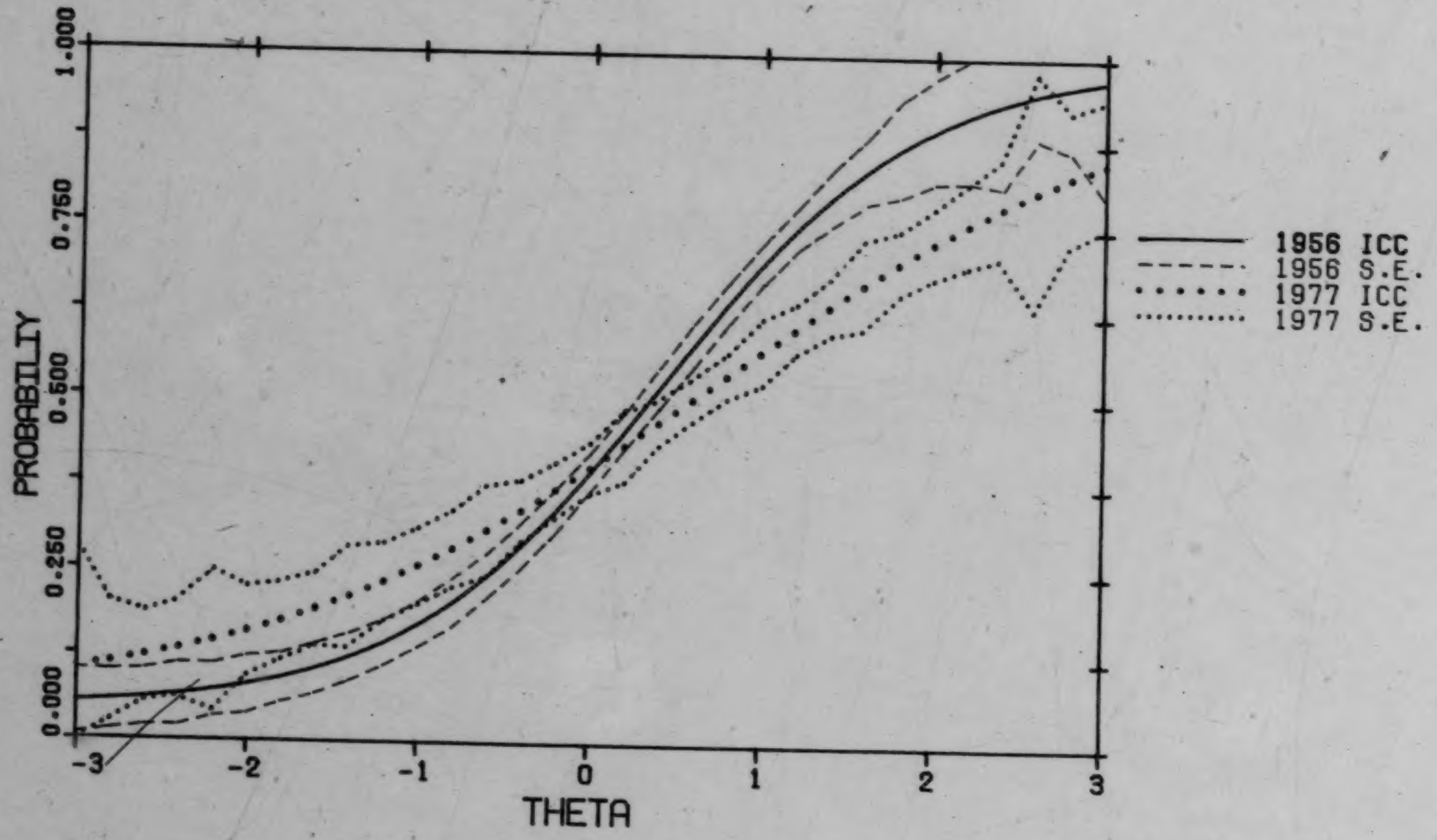


Figure 22

ICC'S FOR ITEM NUMBER 32 : RAVEN

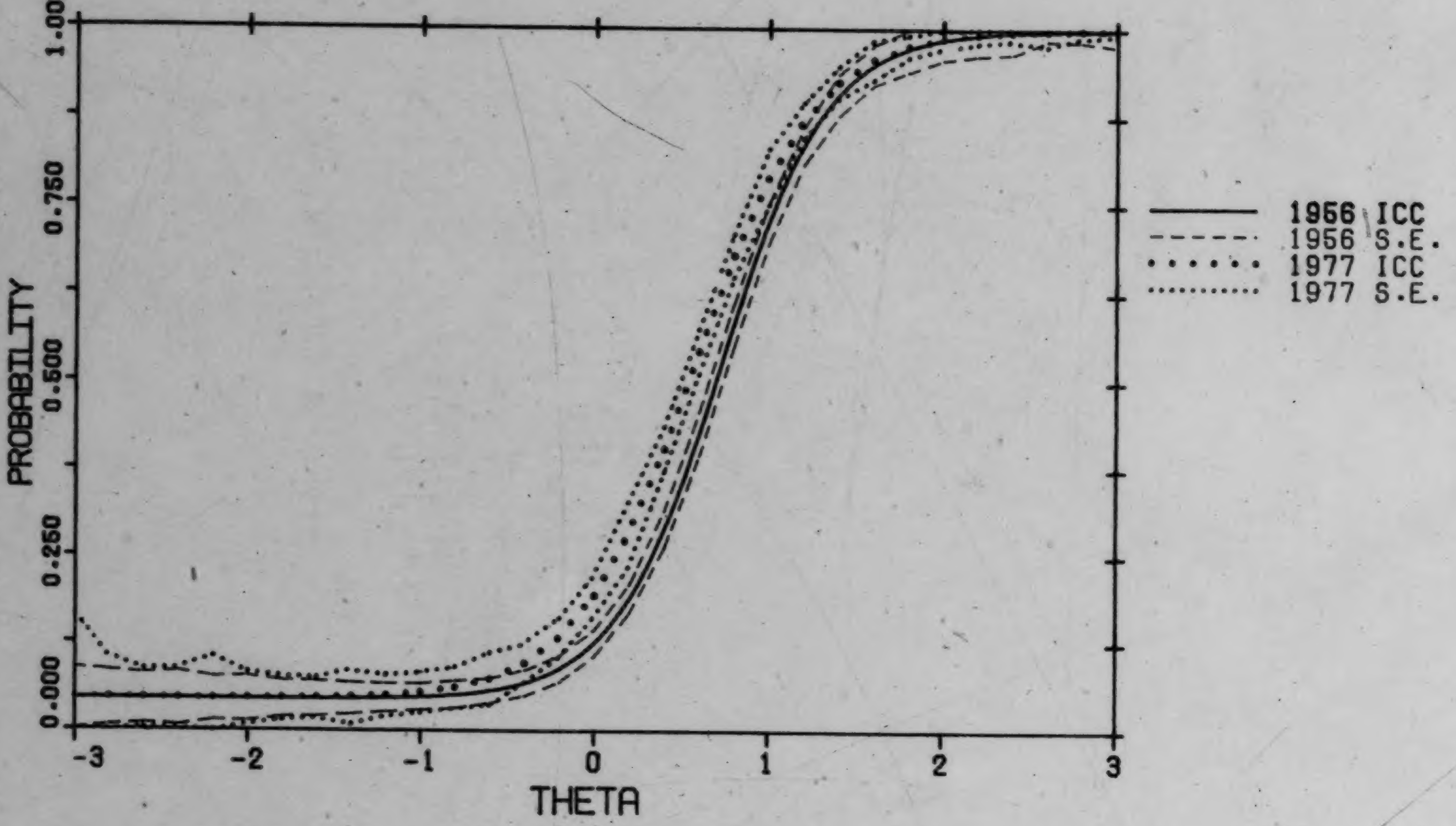


Figure 23

ICC'S FOR ITEM NUMBER 33 : RAVEN

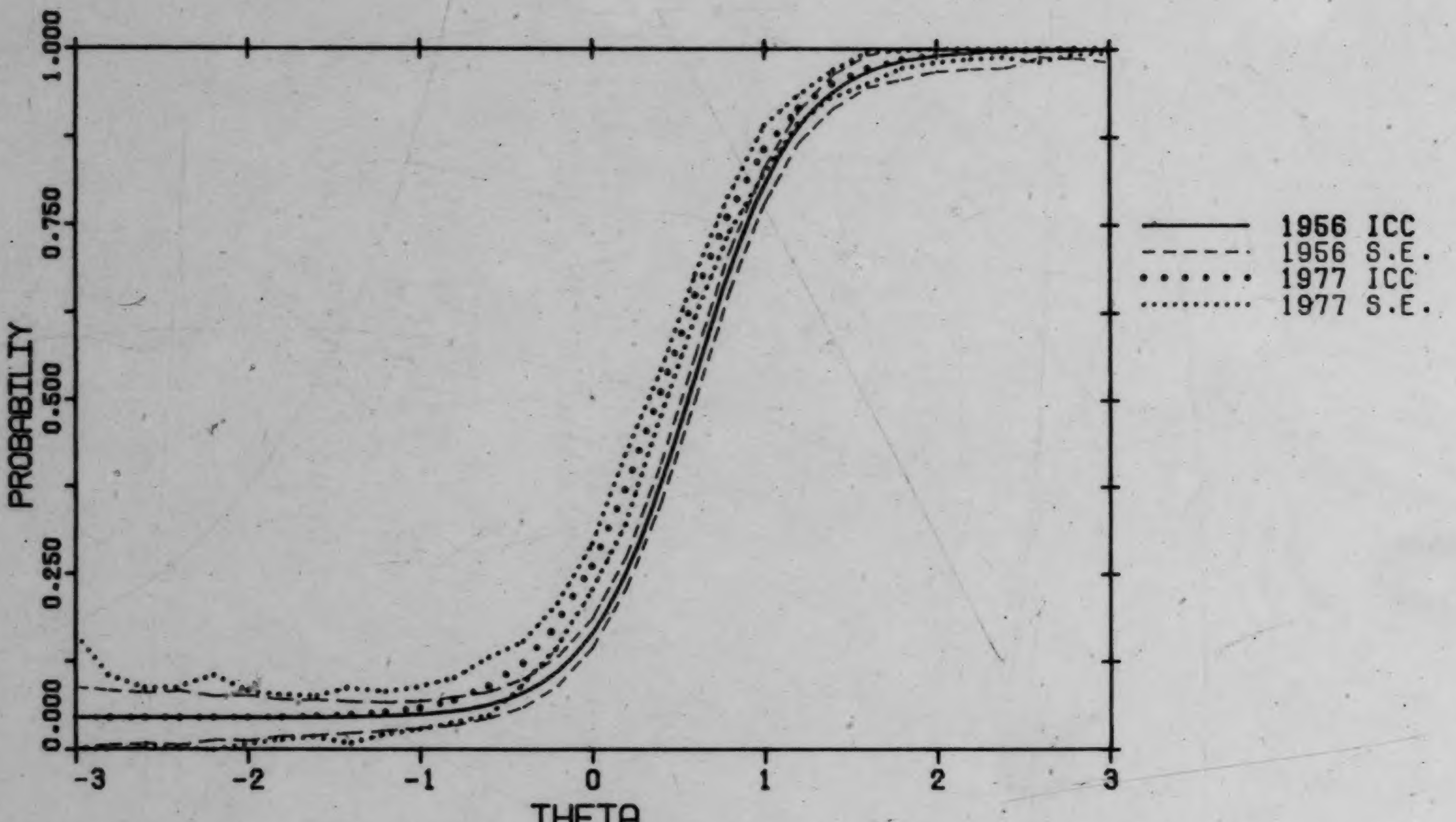


Figure 24

ICC'S FOR ITEM NUMBER 34 : RAVEN

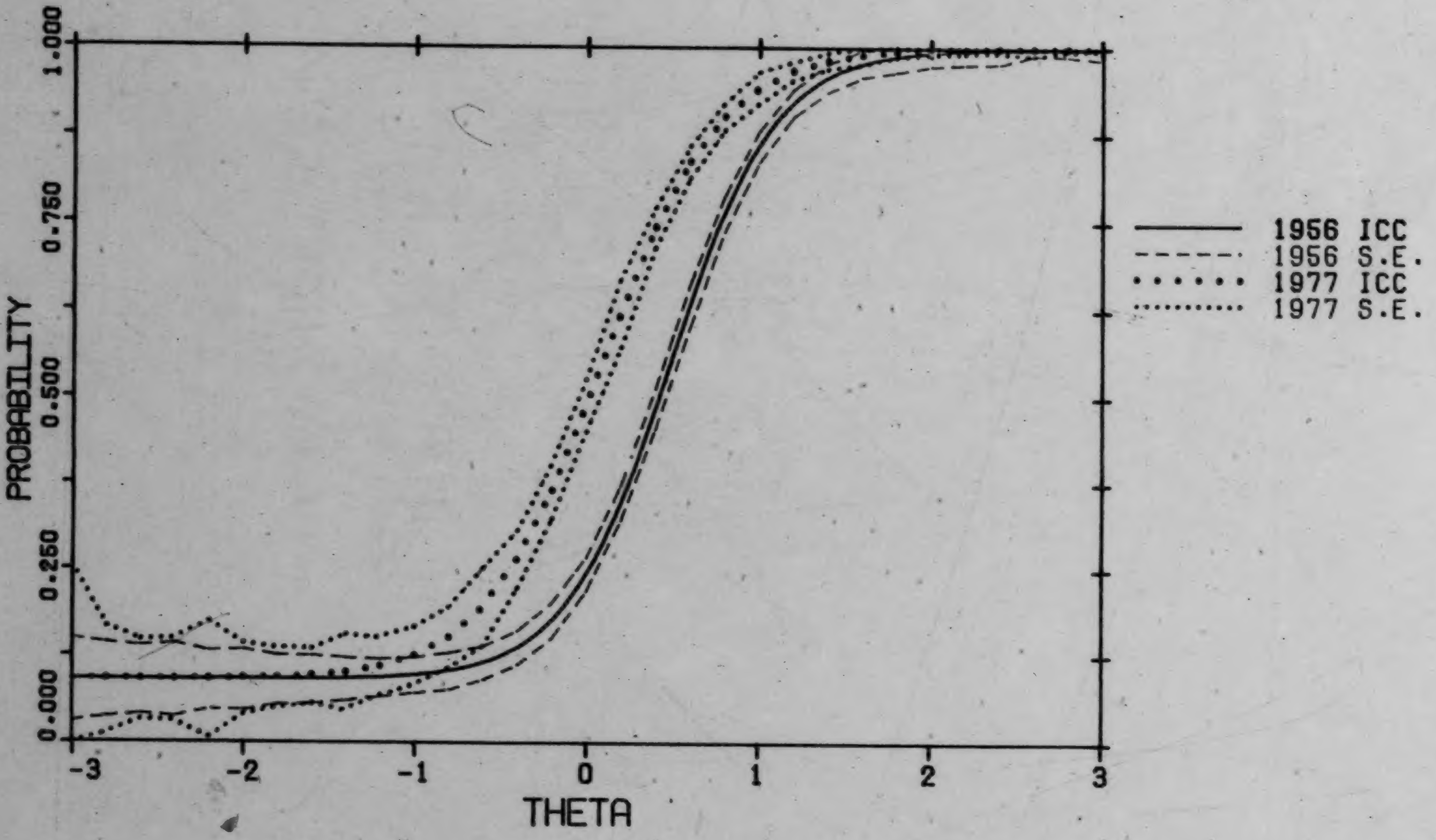


Figure 25

ICC'S FOR ITEM NUMBER 35 : RAVEN

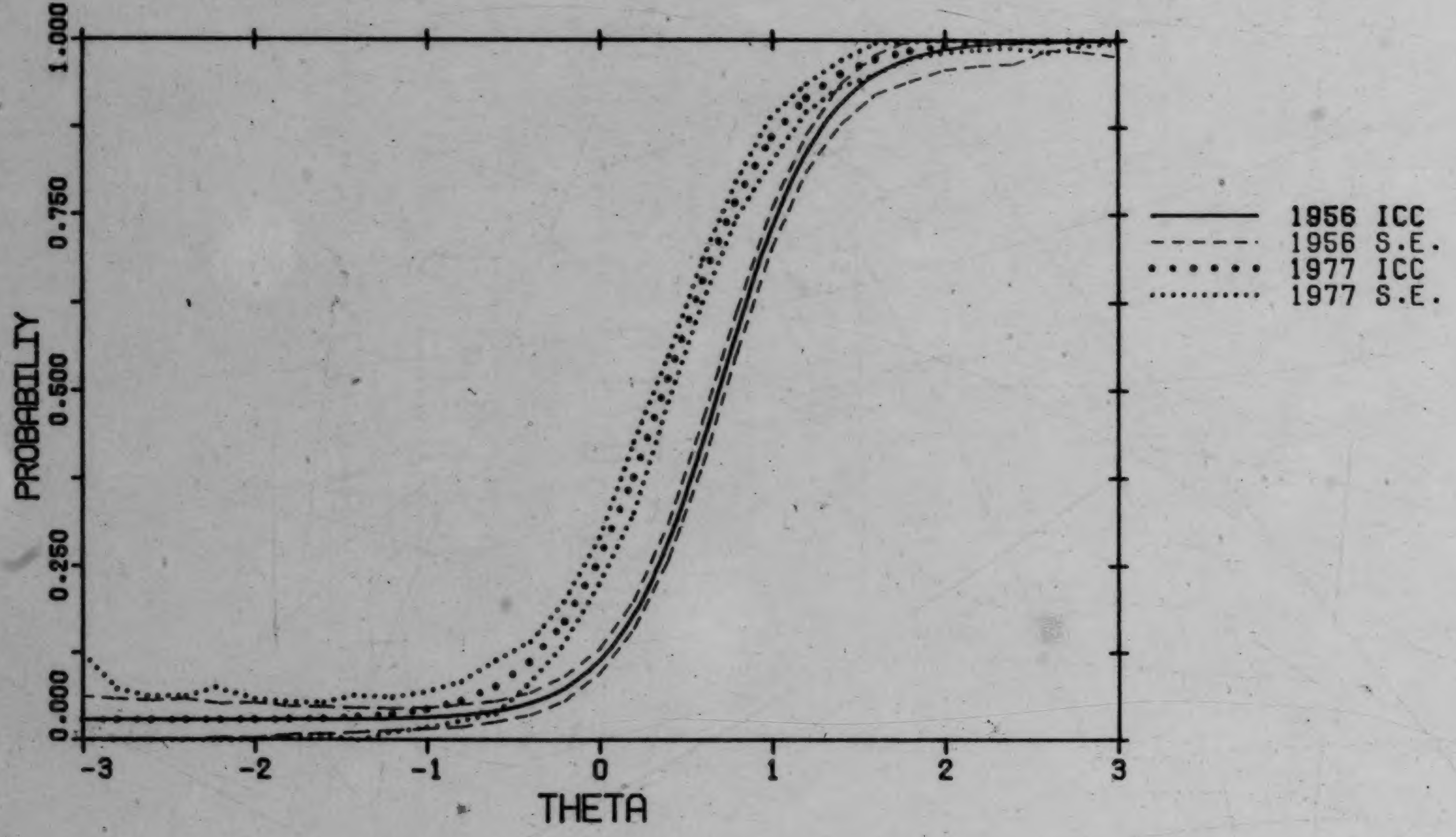


Figure 26

ICC'S FOR ITEM NUMBER 36 : RAVEN

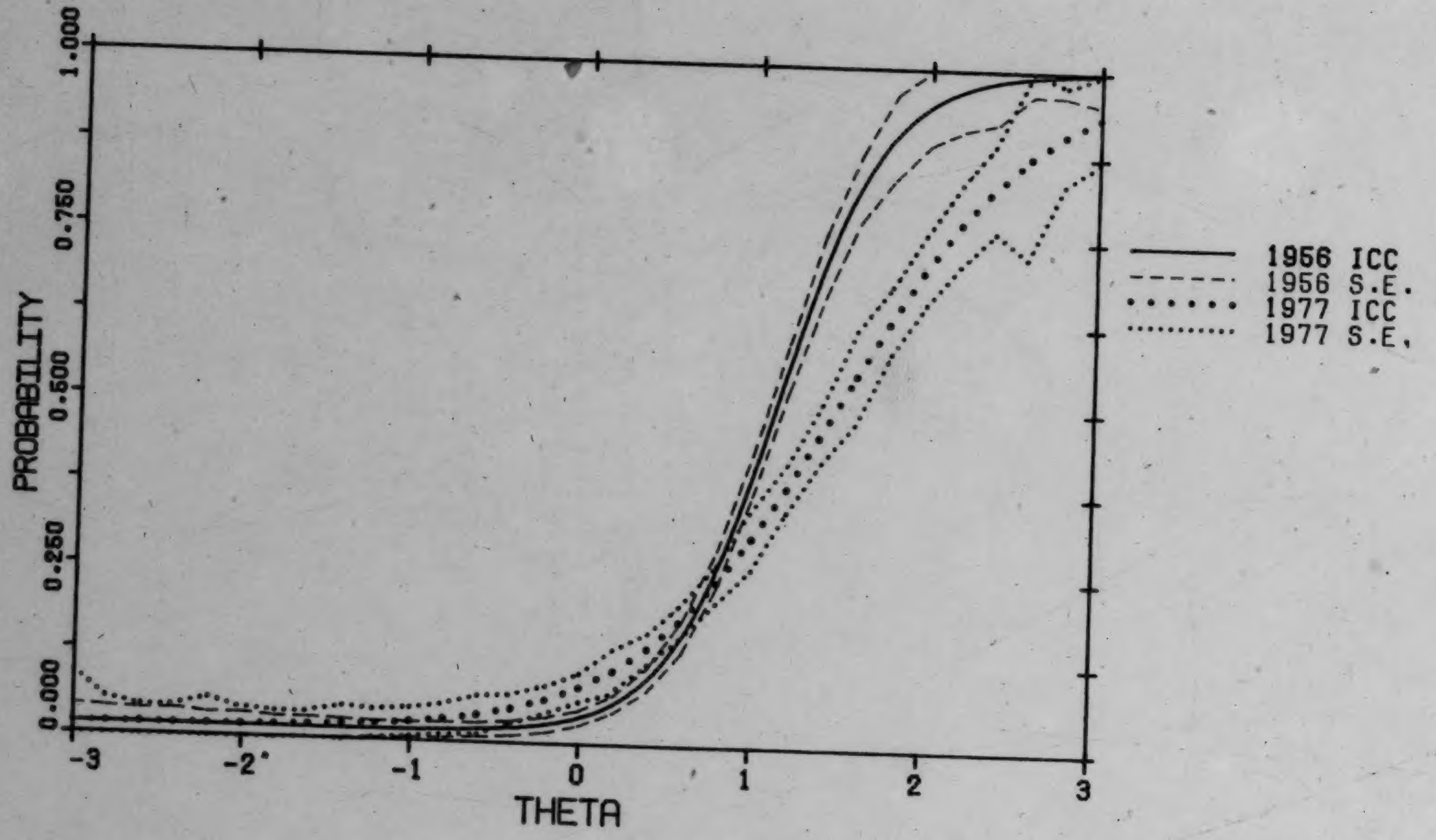


Figure 27

ICC'S FOR ITEM NUMBER 7 : GPR

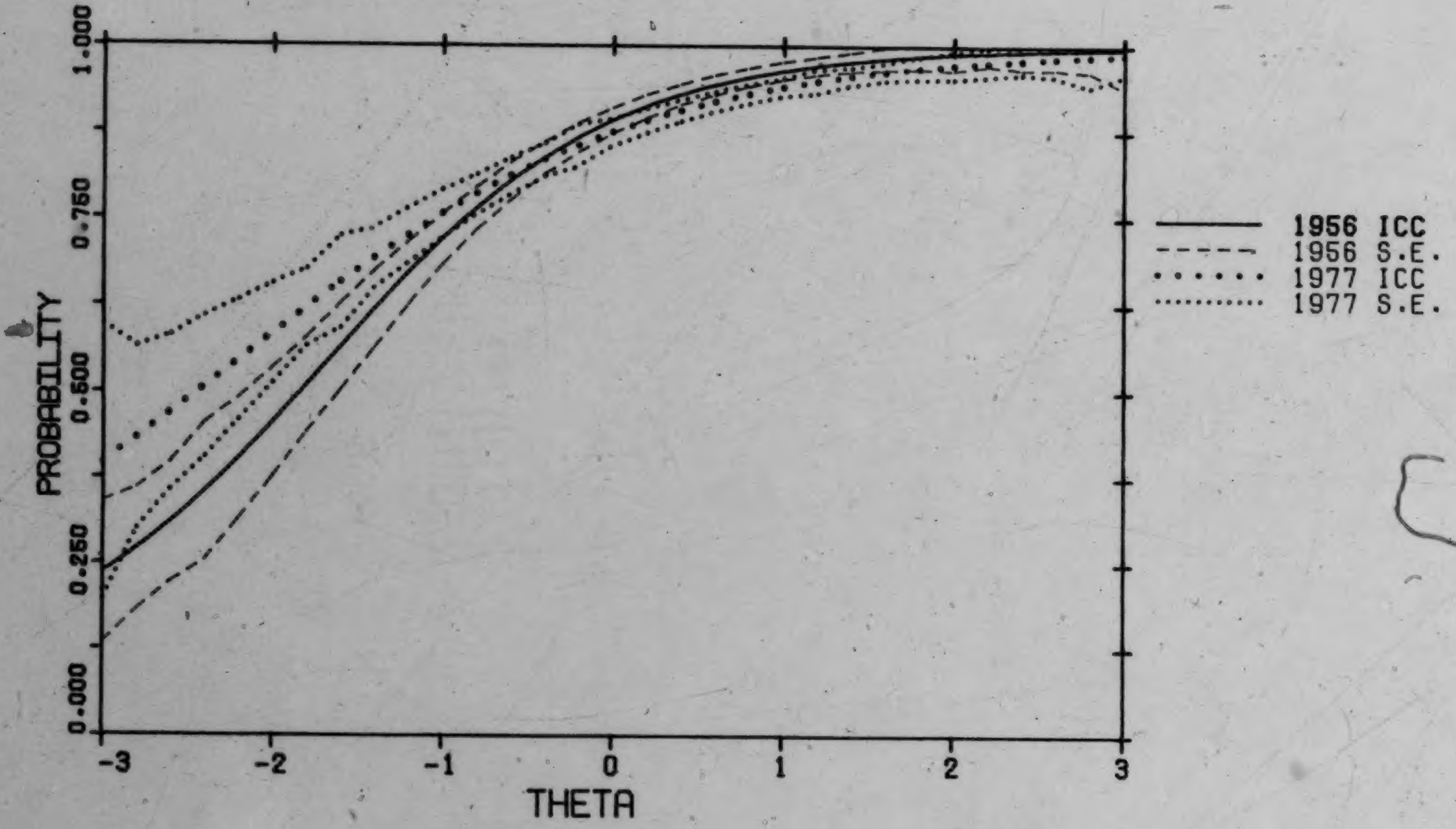


Figure 28

ICC'S FOR ITEM NUMBER 8 : GPR

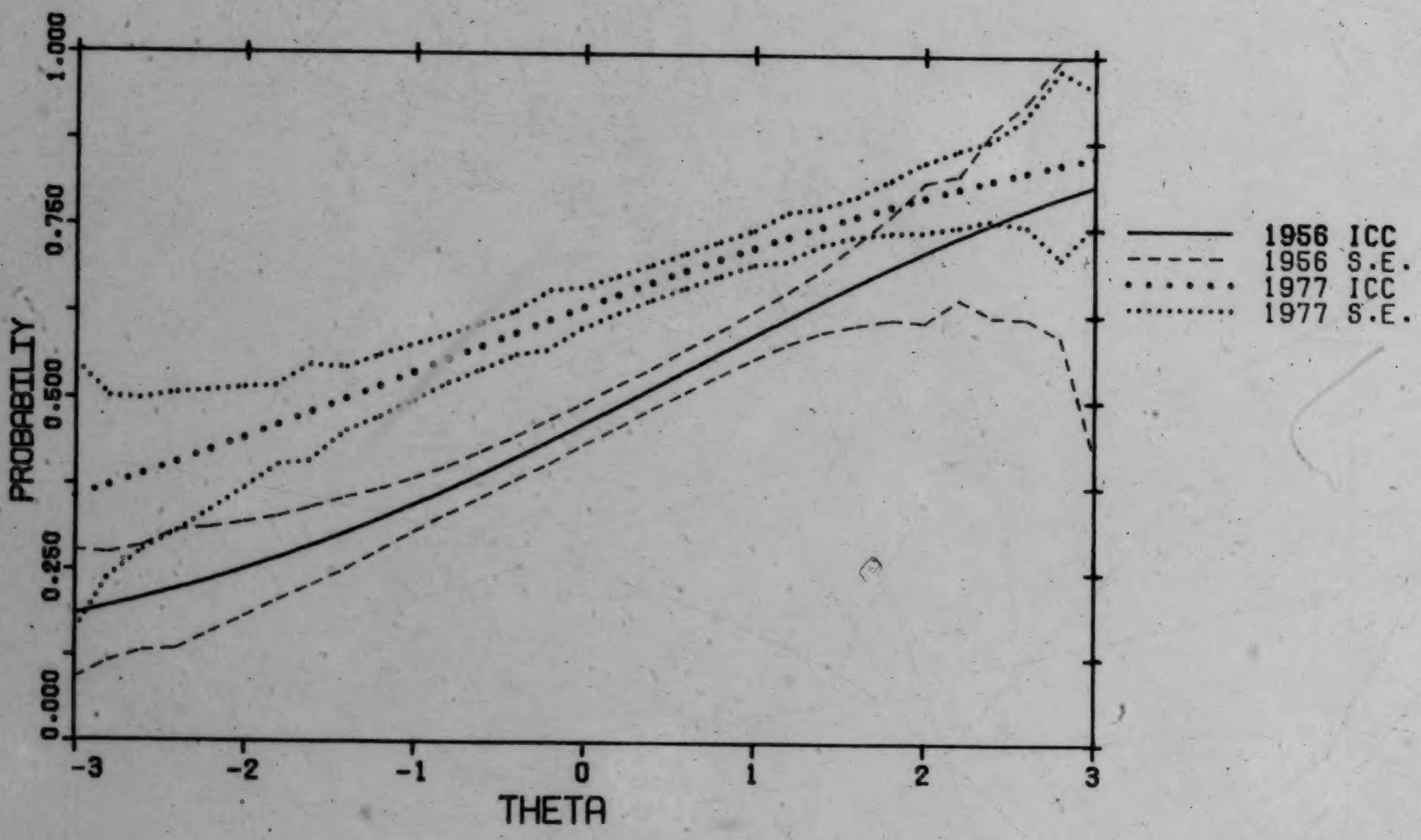


Figure 29

ICC'S FOR ITEM NUMBER 10 : GPR

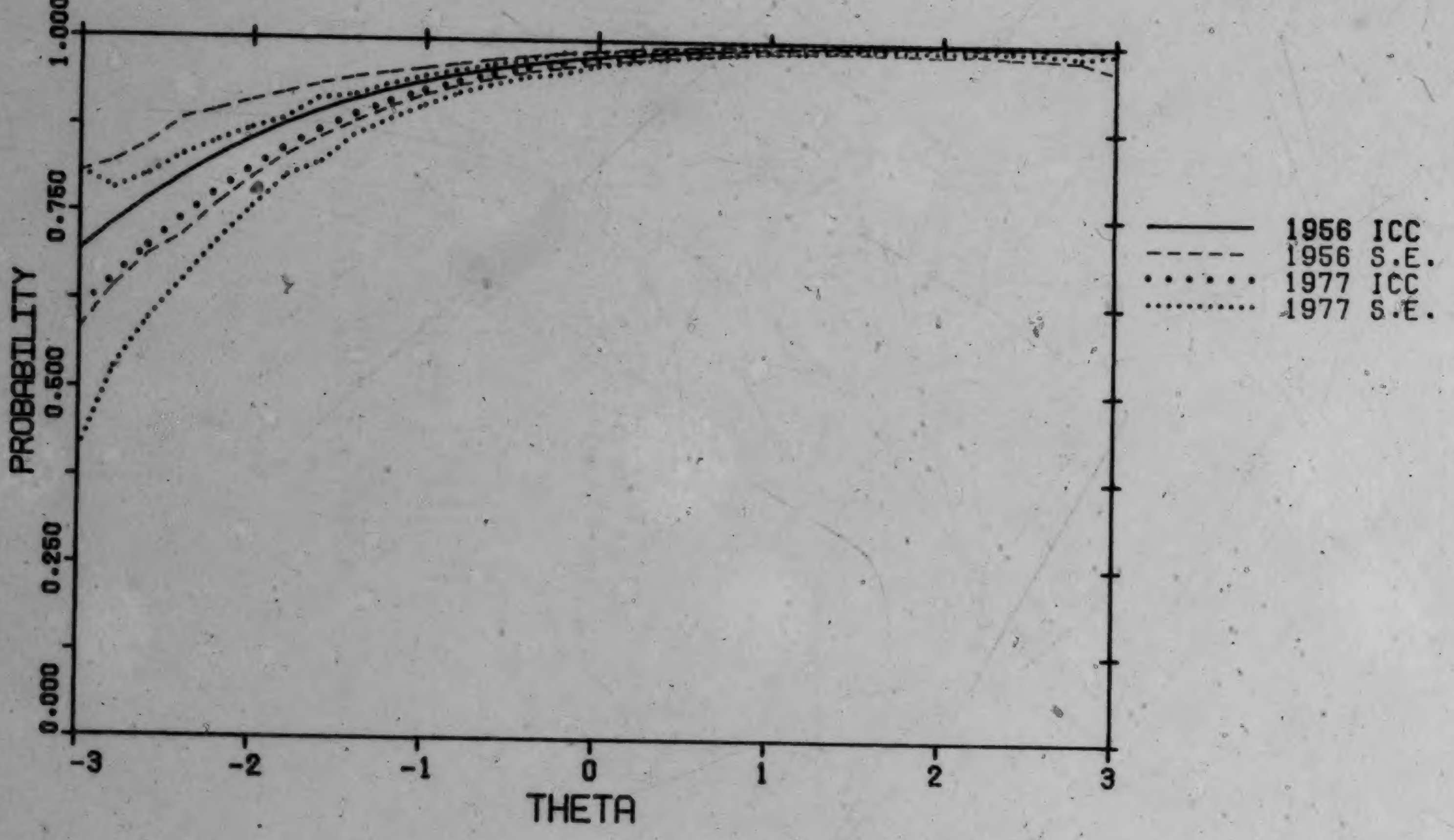


Figure 30

ICC'S FOR ITEM NUMBER 12 : GPR

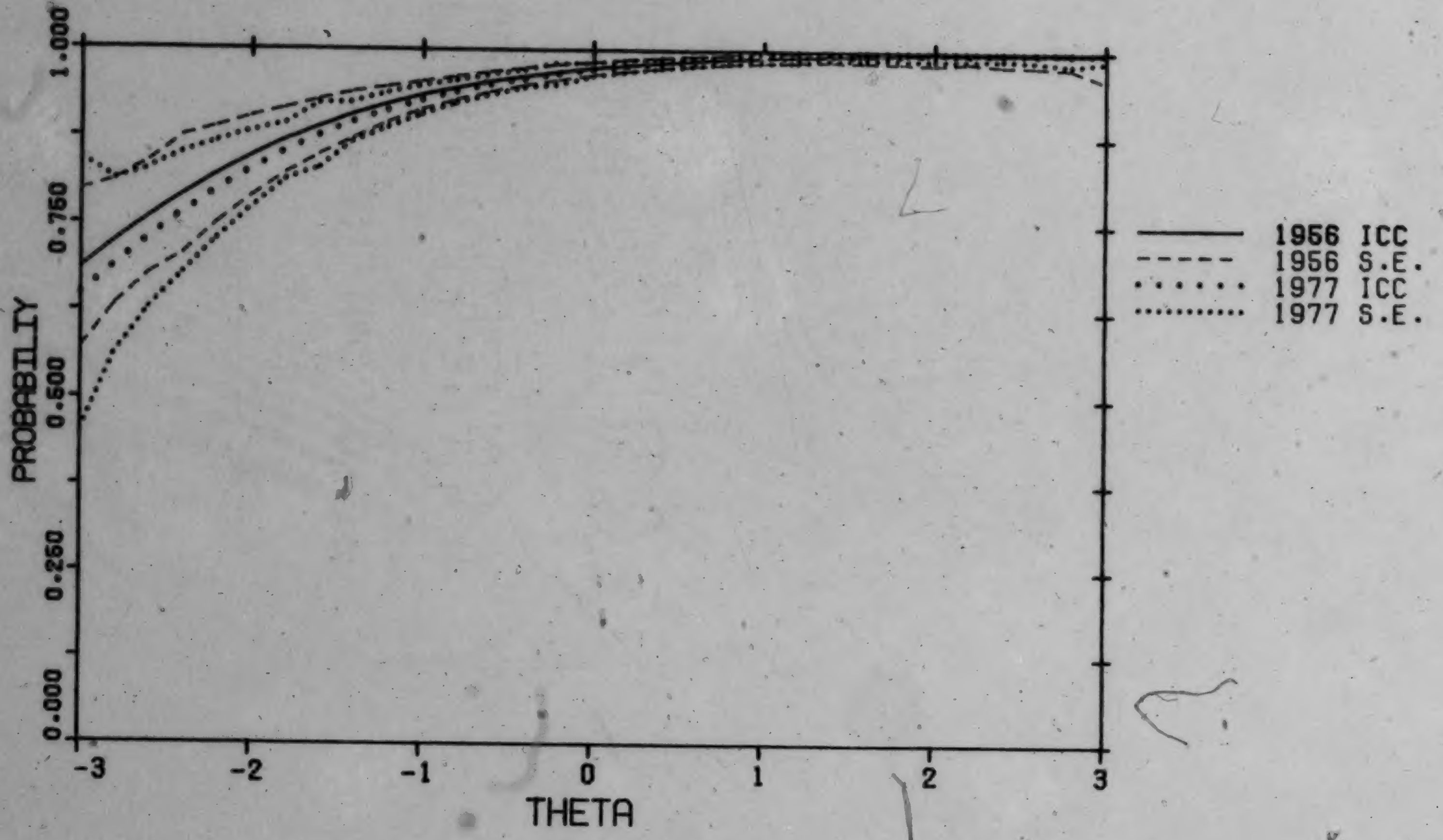


Figure 31

ICC'S FOR ITEM NUMBER 13 : GPR

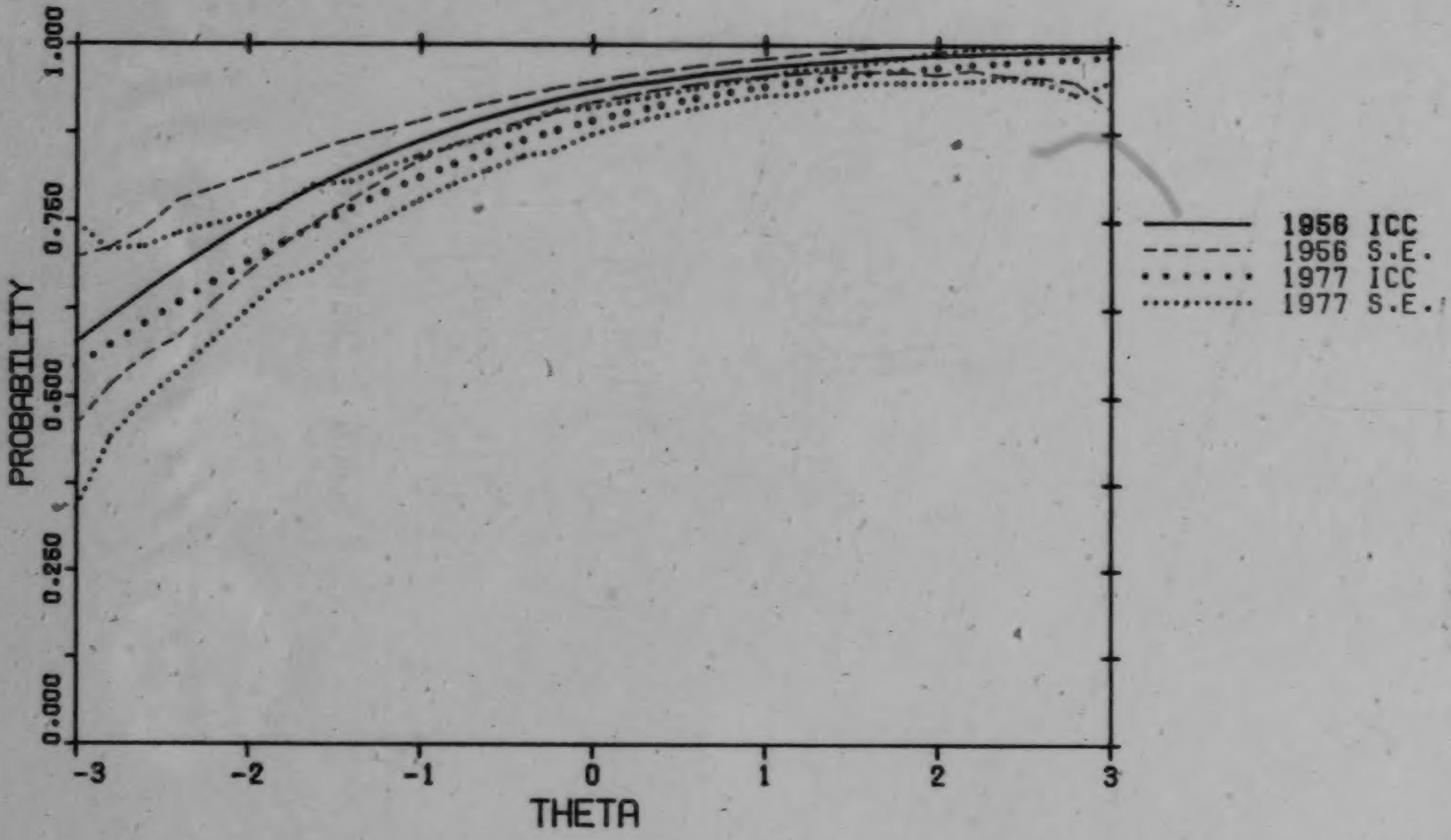


Figure 32

ICC'S FOR ITEM NUMBER 14 : GPR

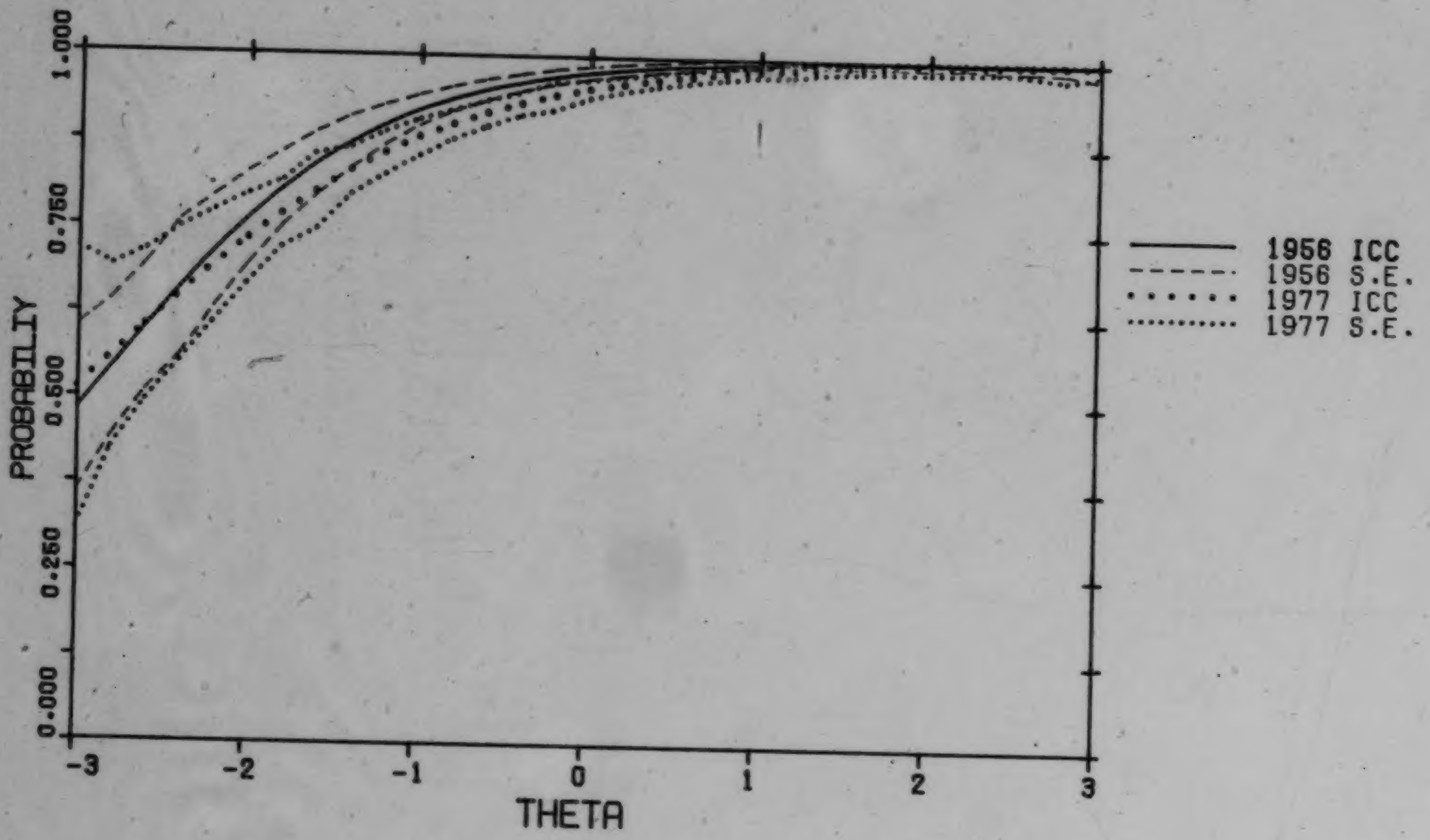


Figure 33

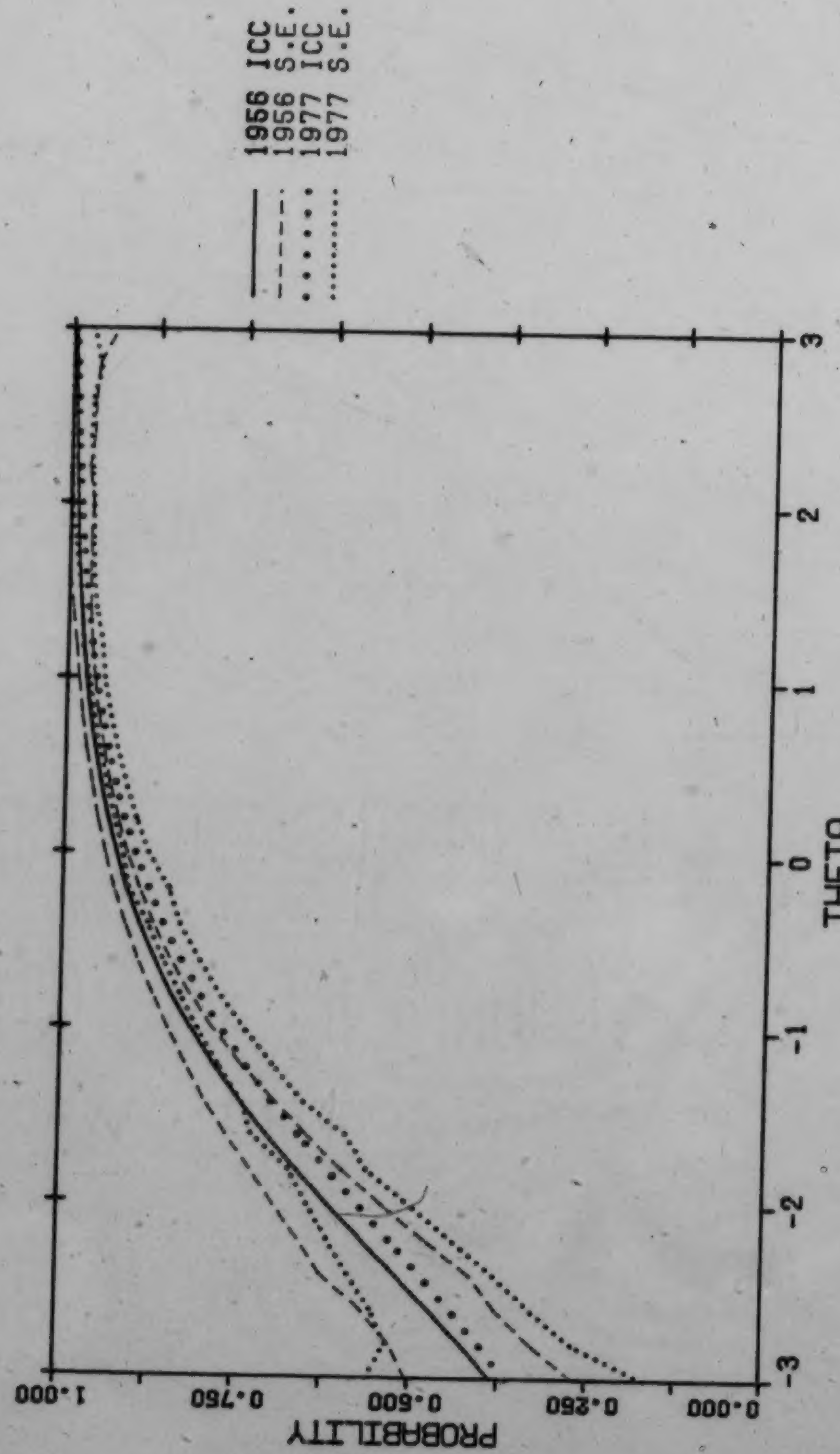

ICC'S FOR ITEM NUMBER 15 : GPR
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 34

ICC'S FOR ITEM NUMBER 16 : GPR

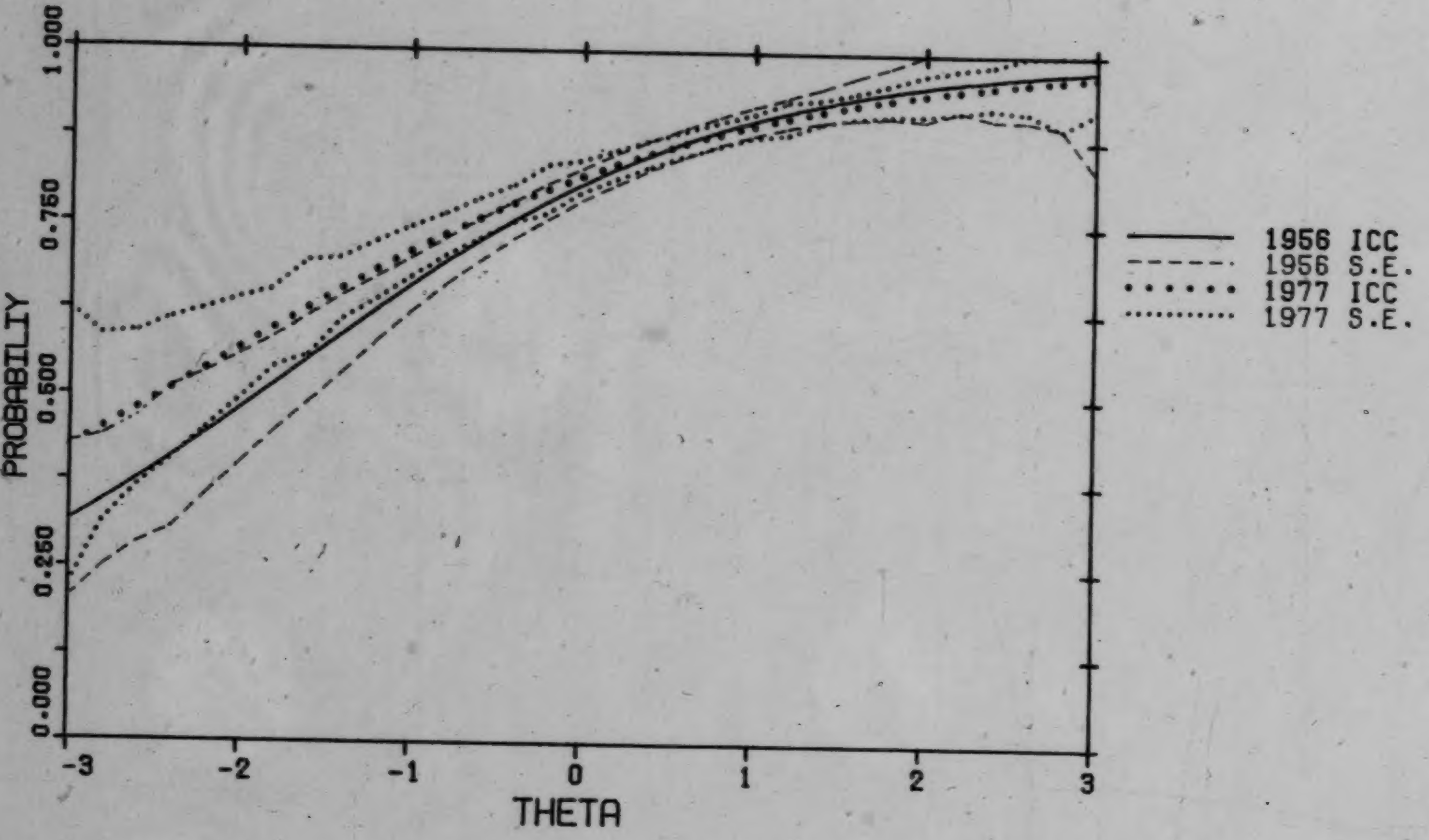


Figure 35

ICC'S FOR ITEM NUMBER 17 : GPR

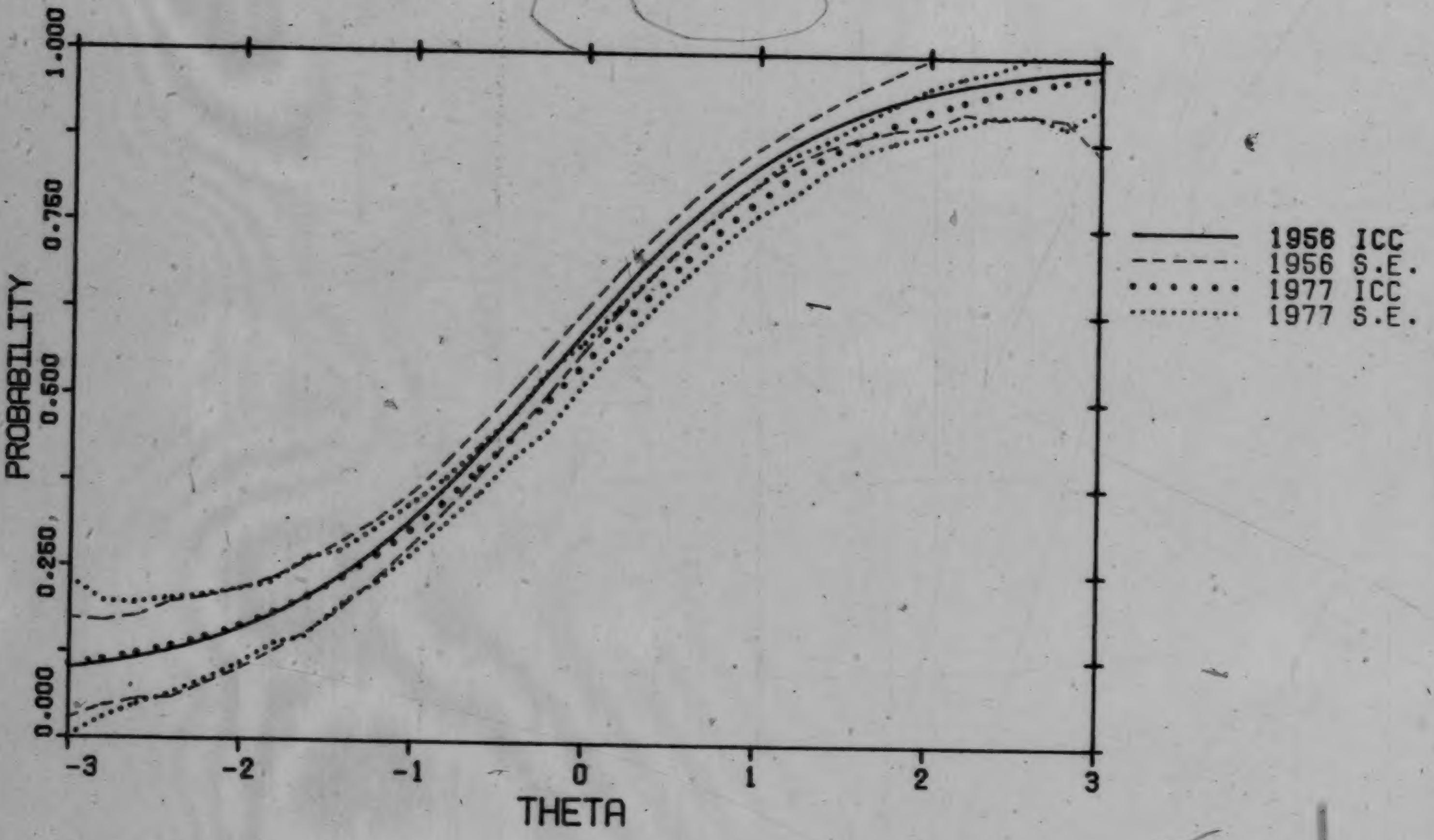


Figure 36

ICC'S FOR ITEM NUMBER 19 : GPR

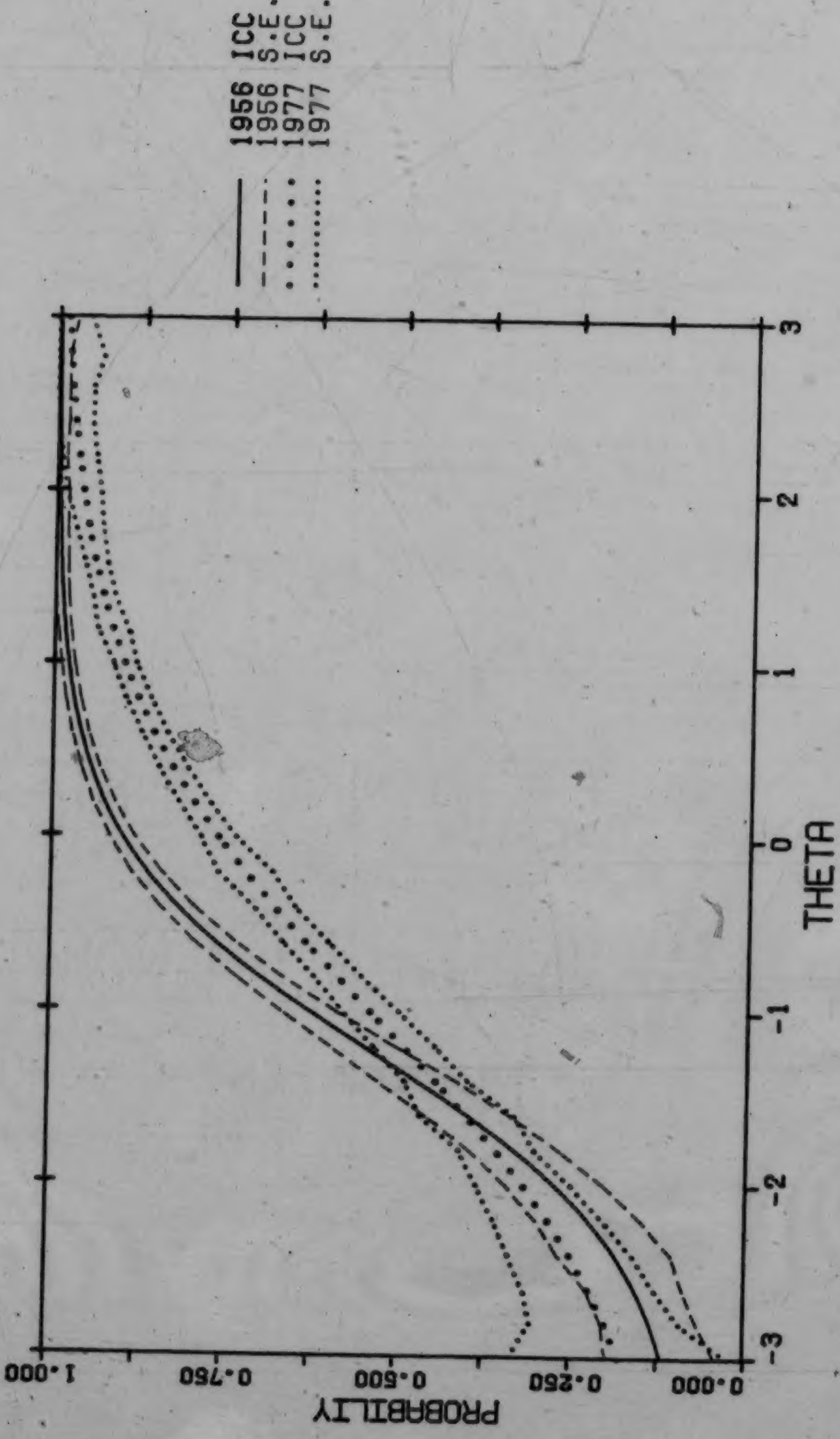


Figure 37

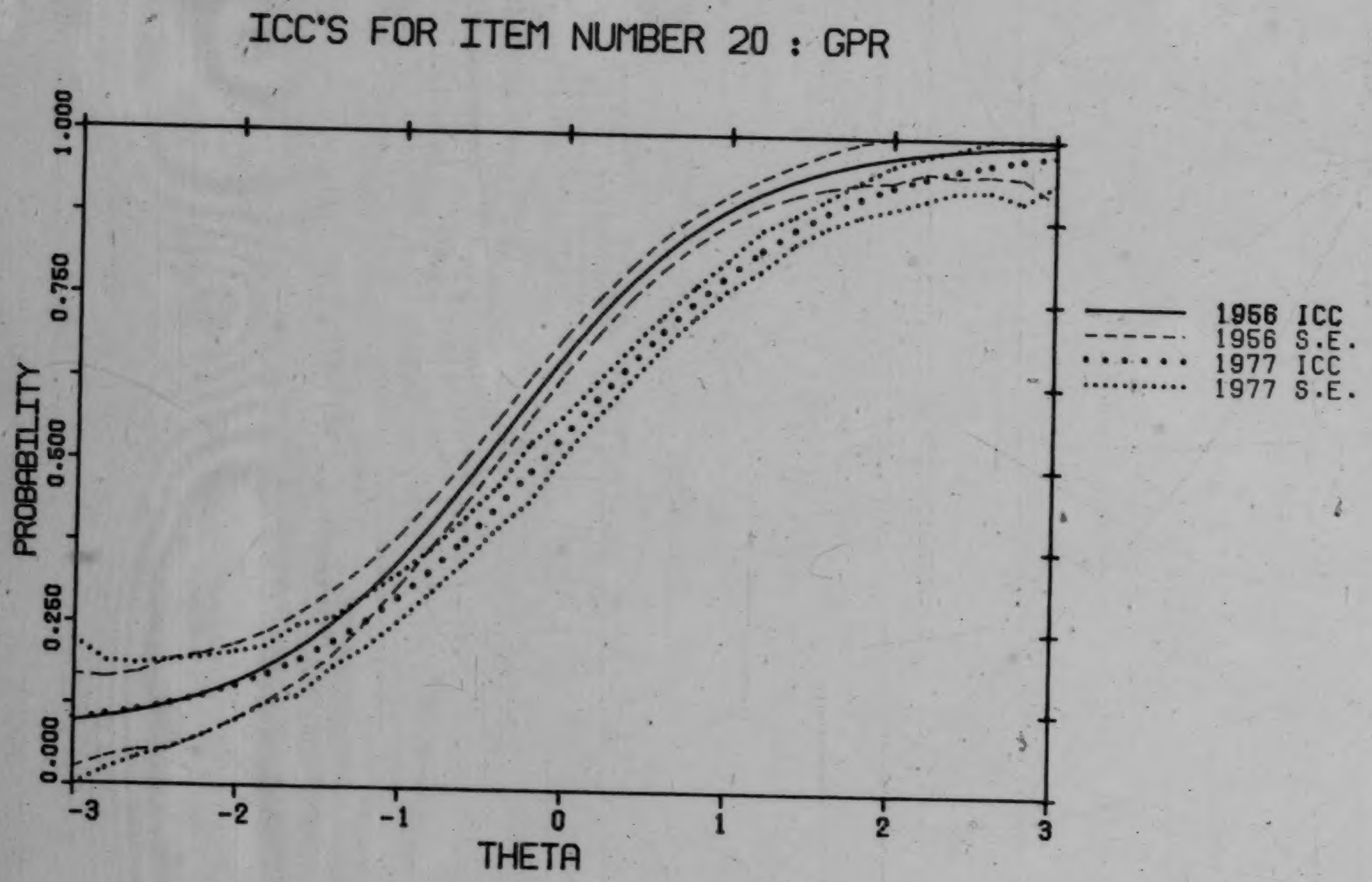

ICC'S FOR ITEM NUMBER 20 : GPR
PROBABILITY
1.000
0.750
0.500
0.250
0.000
-3
-2
-1
0
1
2
3
THETA
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 38

ICC'S FOR ITEM NUMBER 21 : GPR

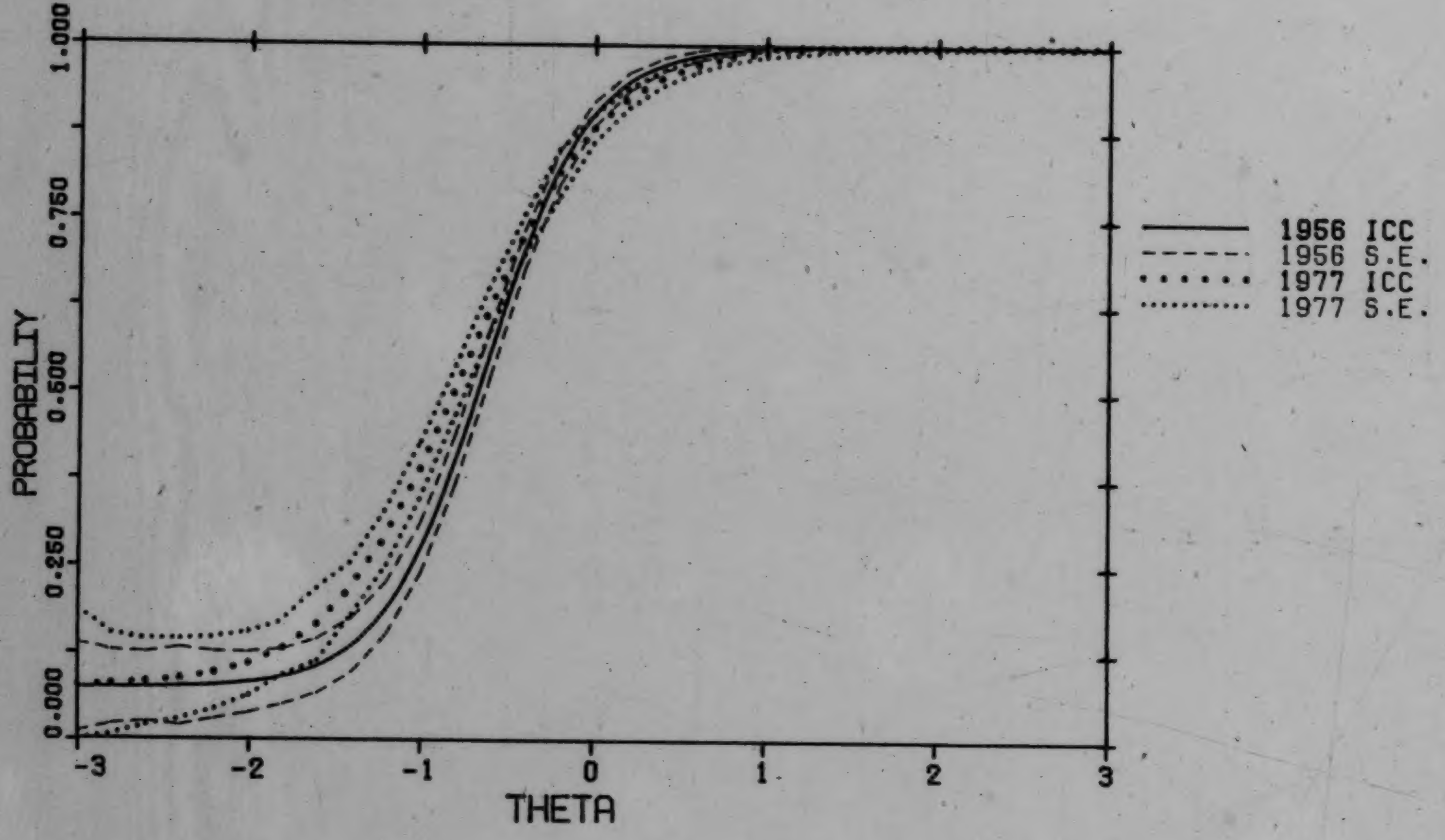


Figure 39

ICC'S FOR ITEM NUMBER 22 : GPR

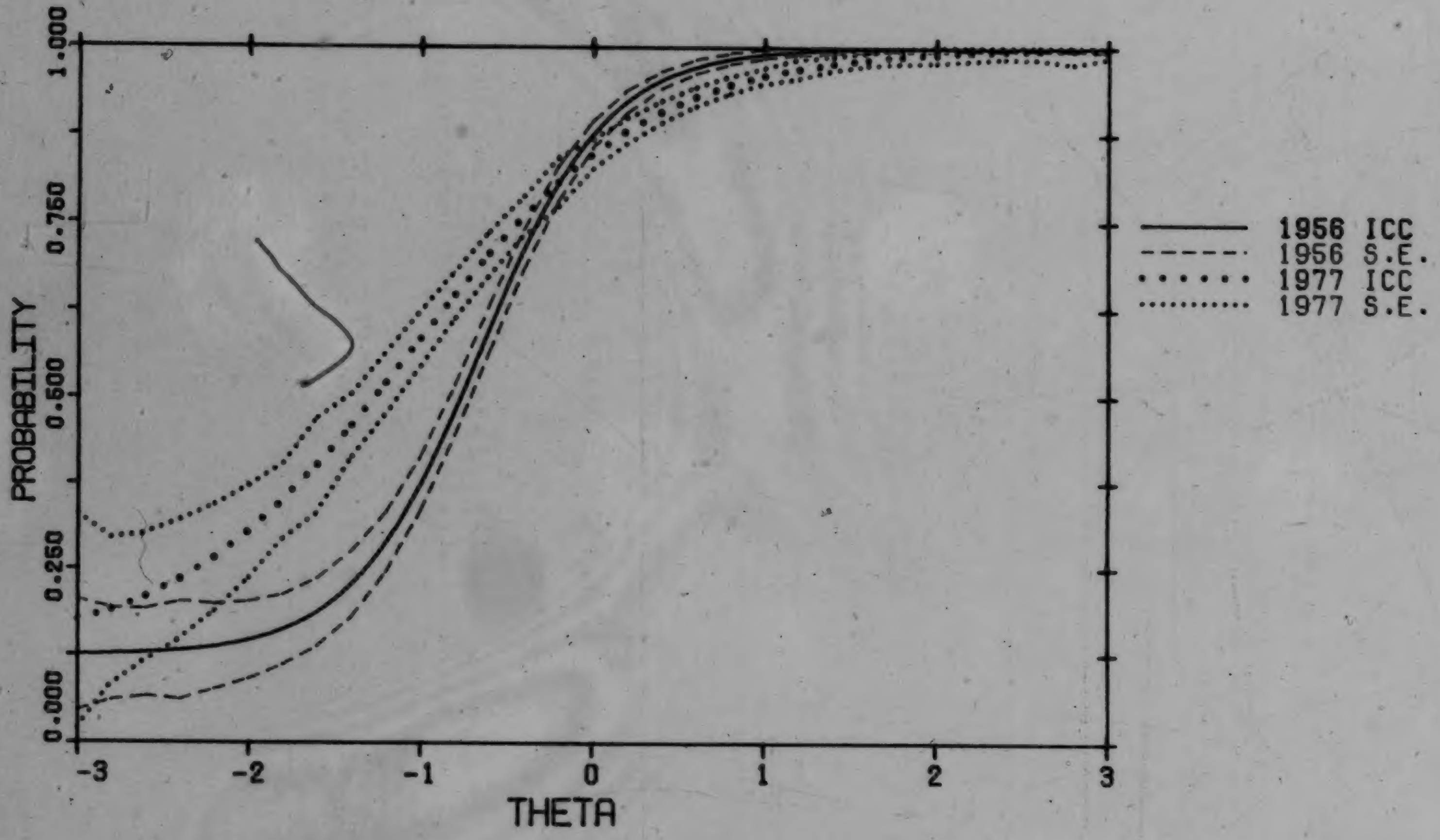


Figure 40

ICC'S FOR ITEM NUMBER 23 : GPR

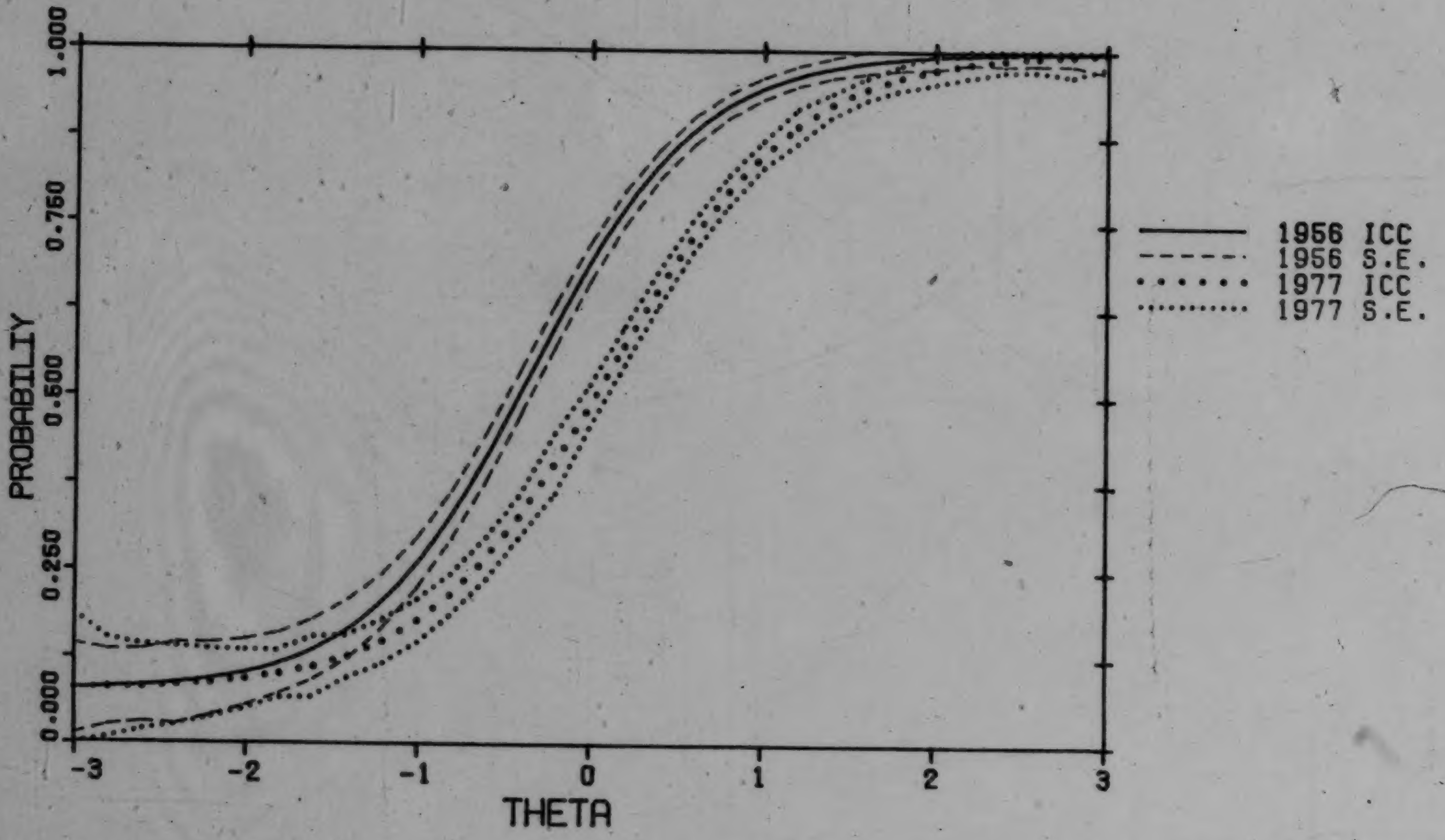


Figure 41

ICC'S FOR ITEM NUMBER 24 : GPR

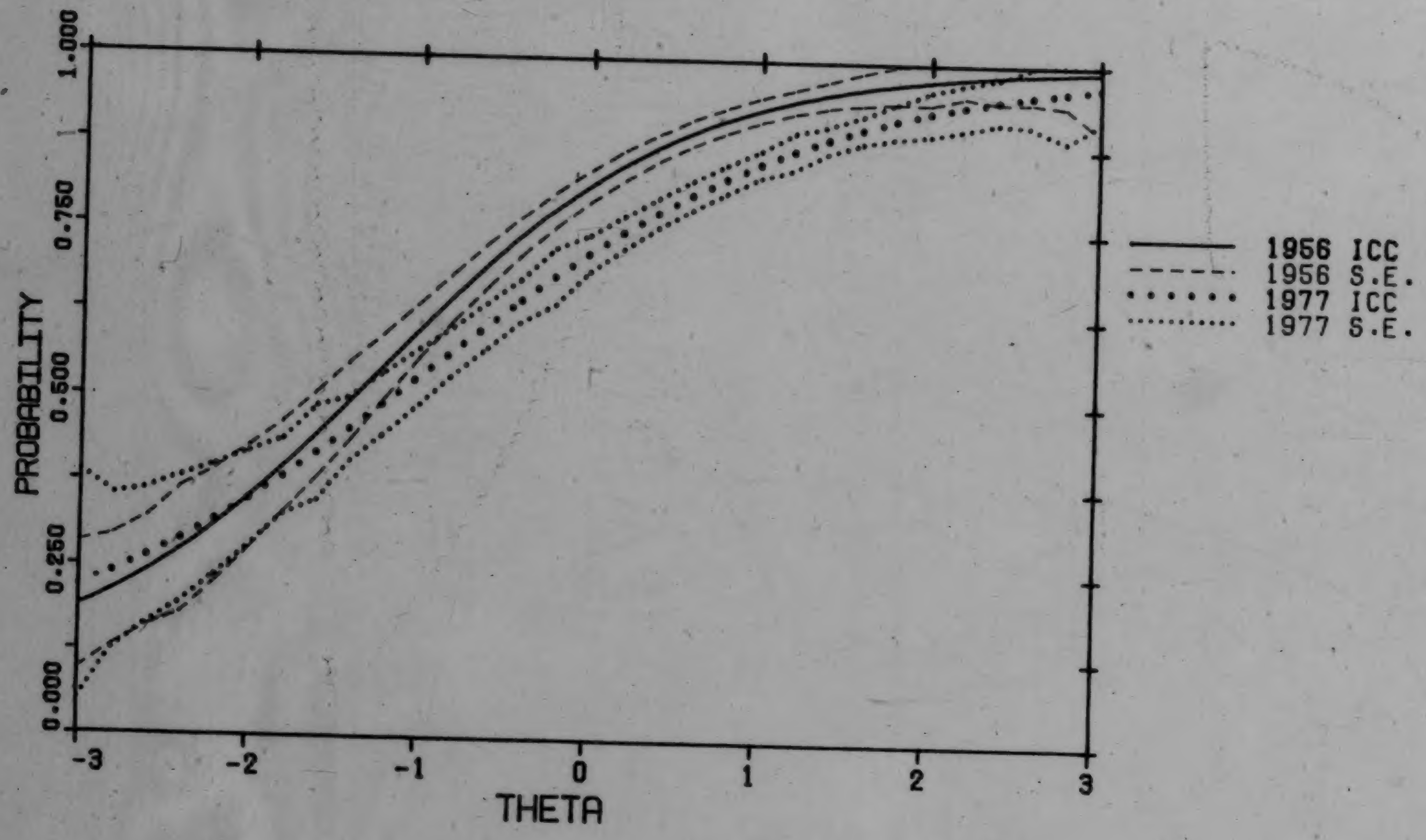


Figure 42

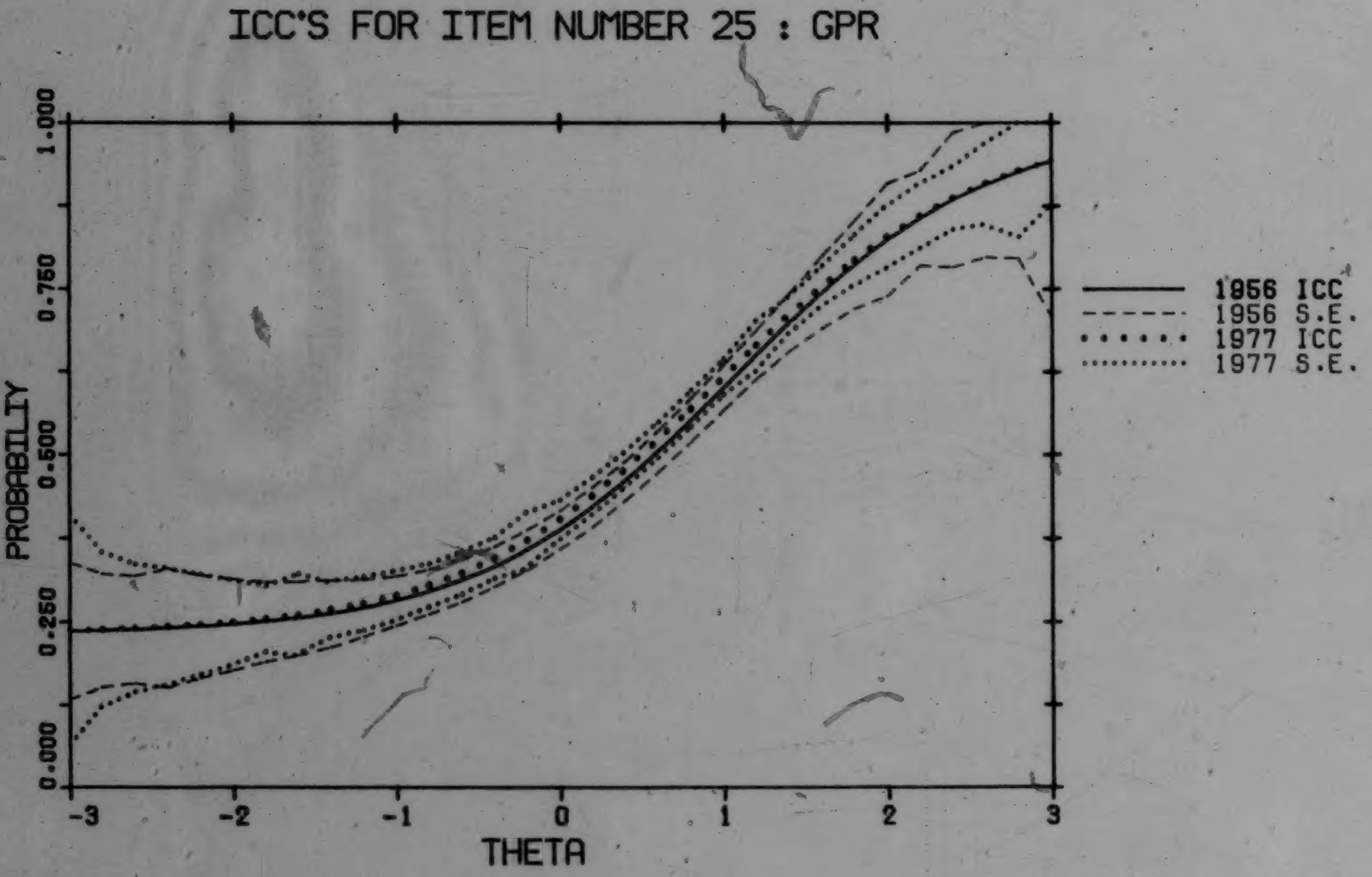

ICC'S FOR ITEM NUMBER 25 : GPR
PROBABILITY
0.000
0.250
0.500
0.750
1.000
-3
-2
-1
0
1
2
3
THETA
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 43

ICC'S FOR ITEM NUMBER 26 : GPR

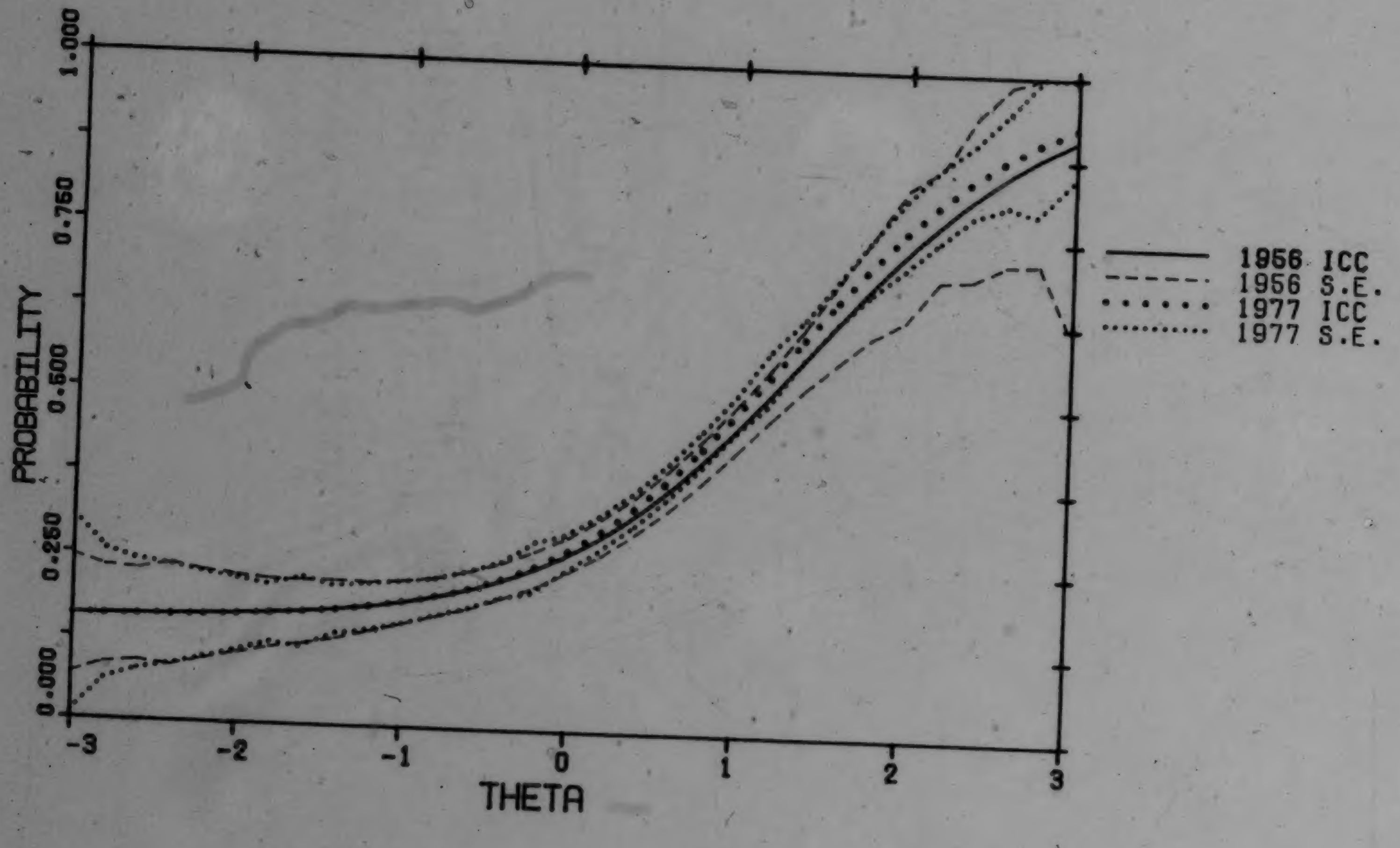


Figure 44

ICC'S FOR ITEM NUMBER 27 : GPR

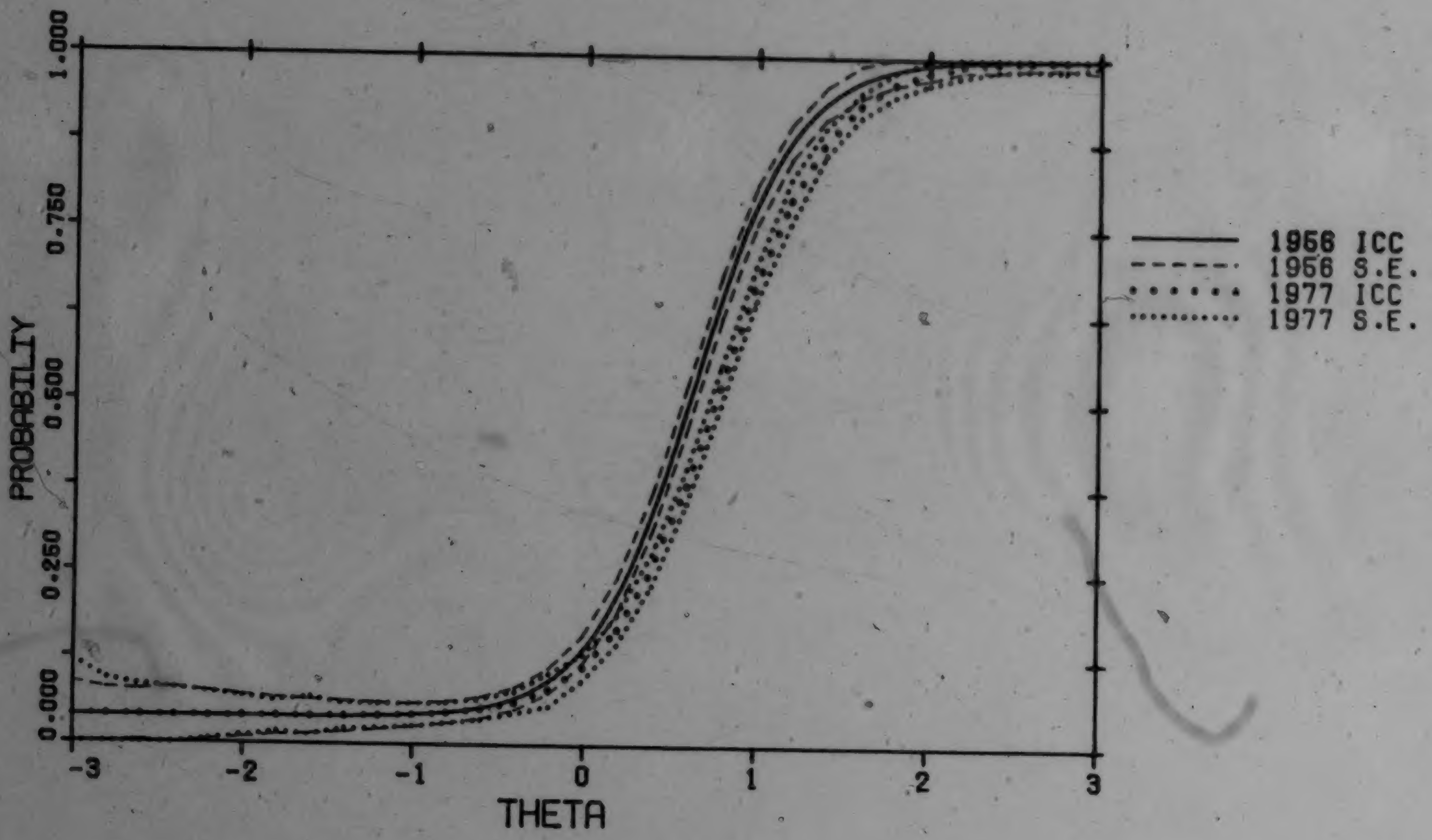


Figure 45

ICC'S FOR ITEM NUMBER 28 : GPR

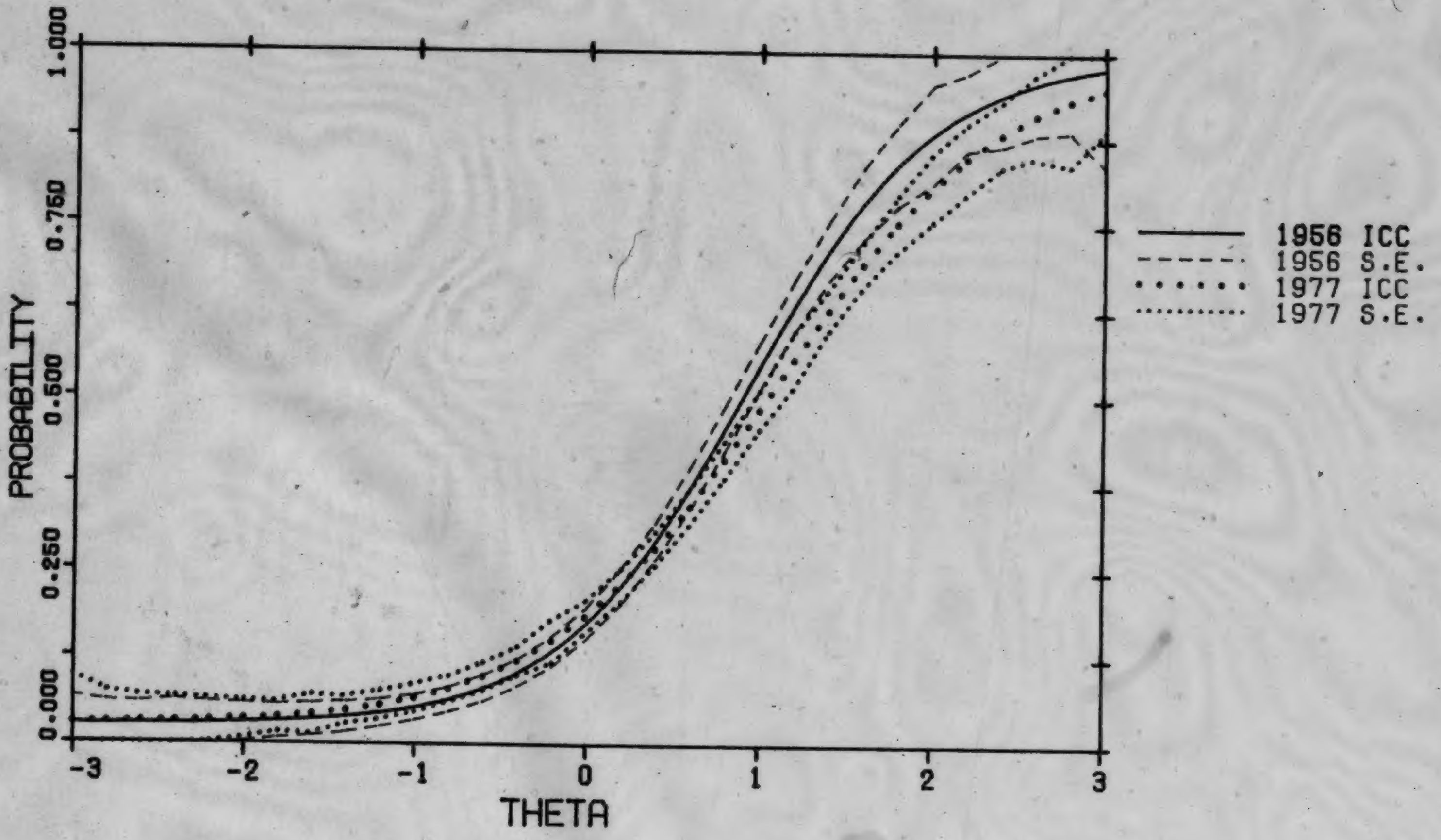


Figure 46

ICC'S FOR ITEM NUMBER 29 : GPR

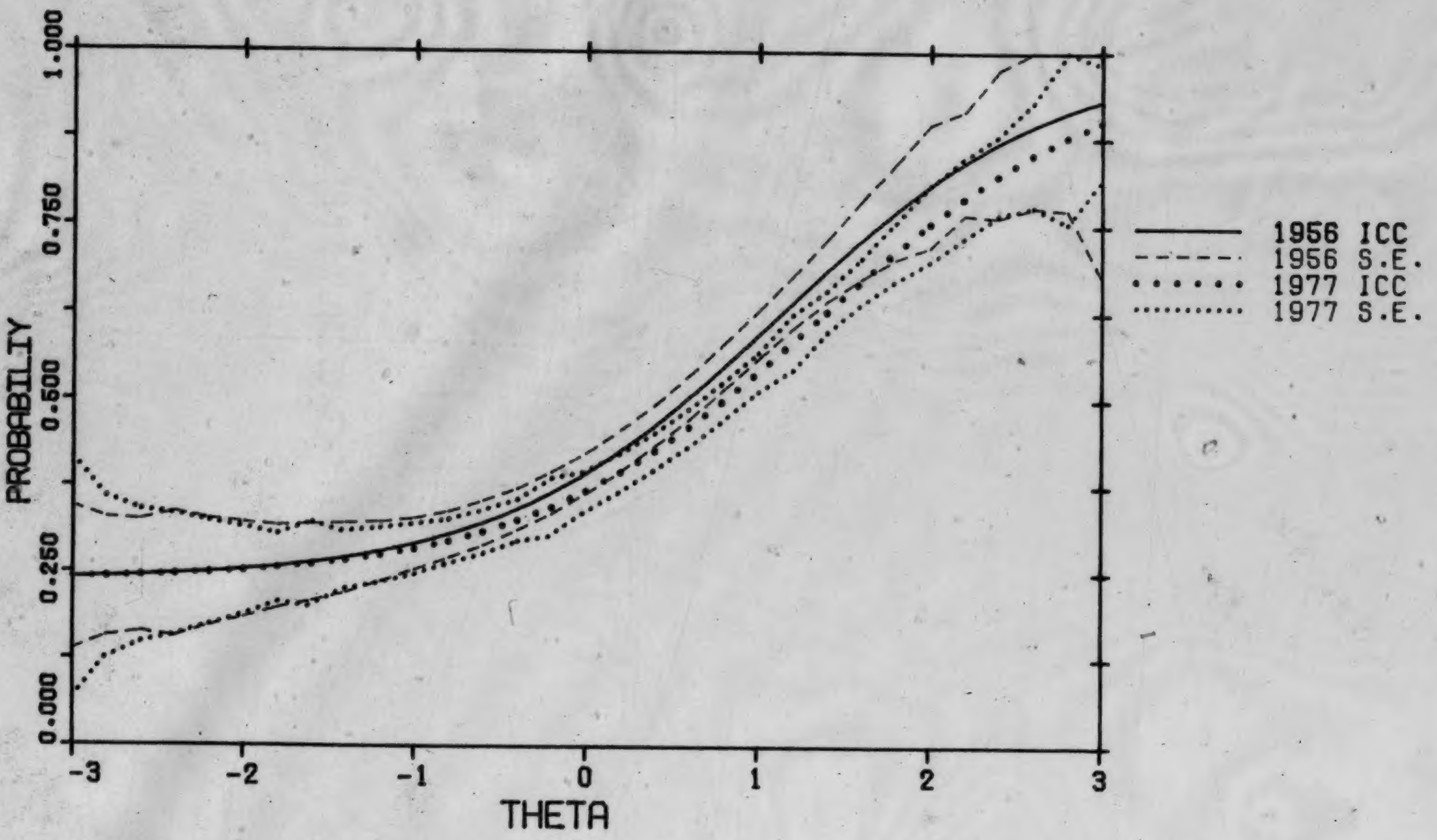


Figure 47

ICC'S FOR ITEM NUMBER 30 : GPR

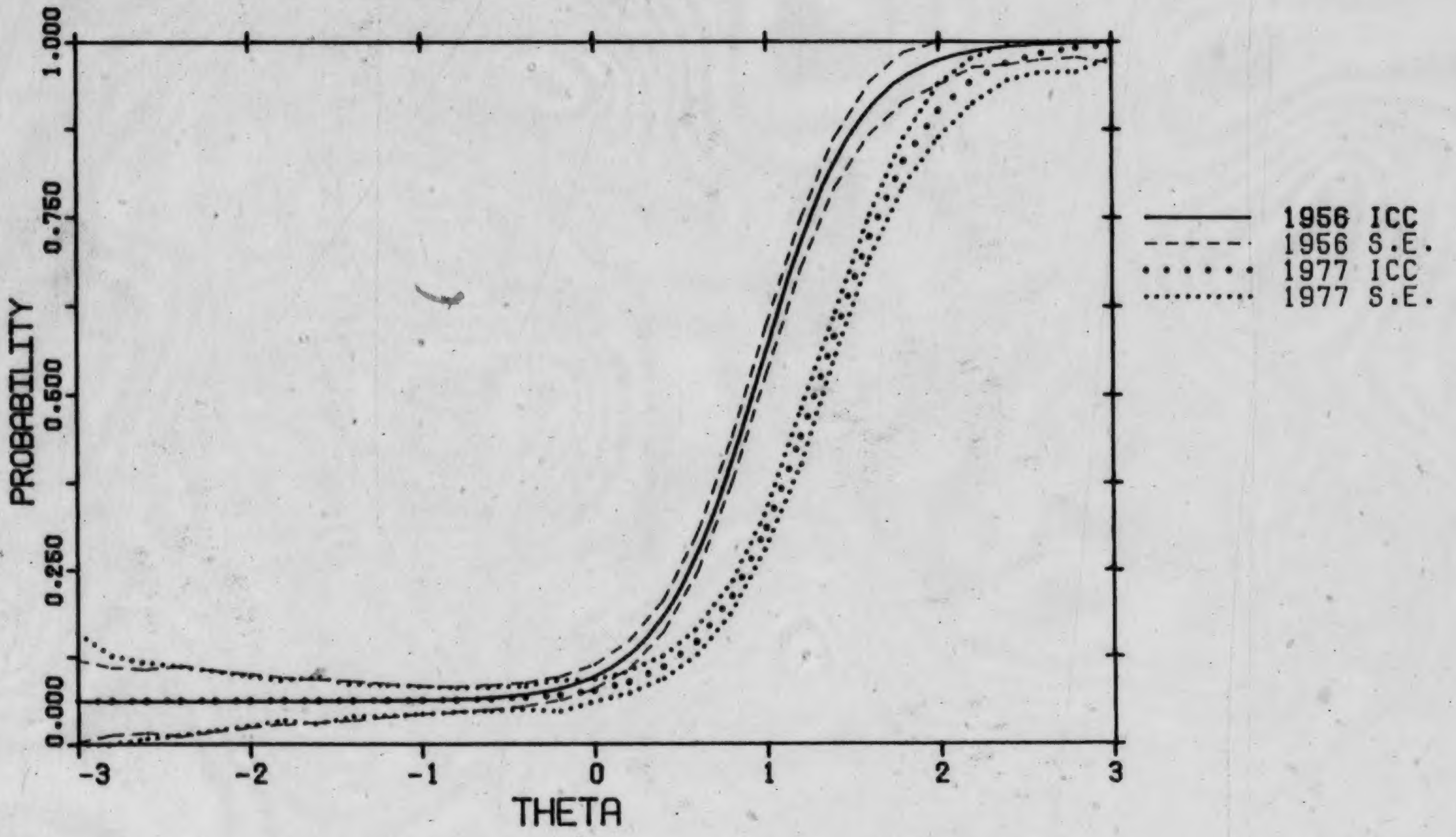


Figure 48

ICC'S FOR ITEM NUMBER 31 : GPR

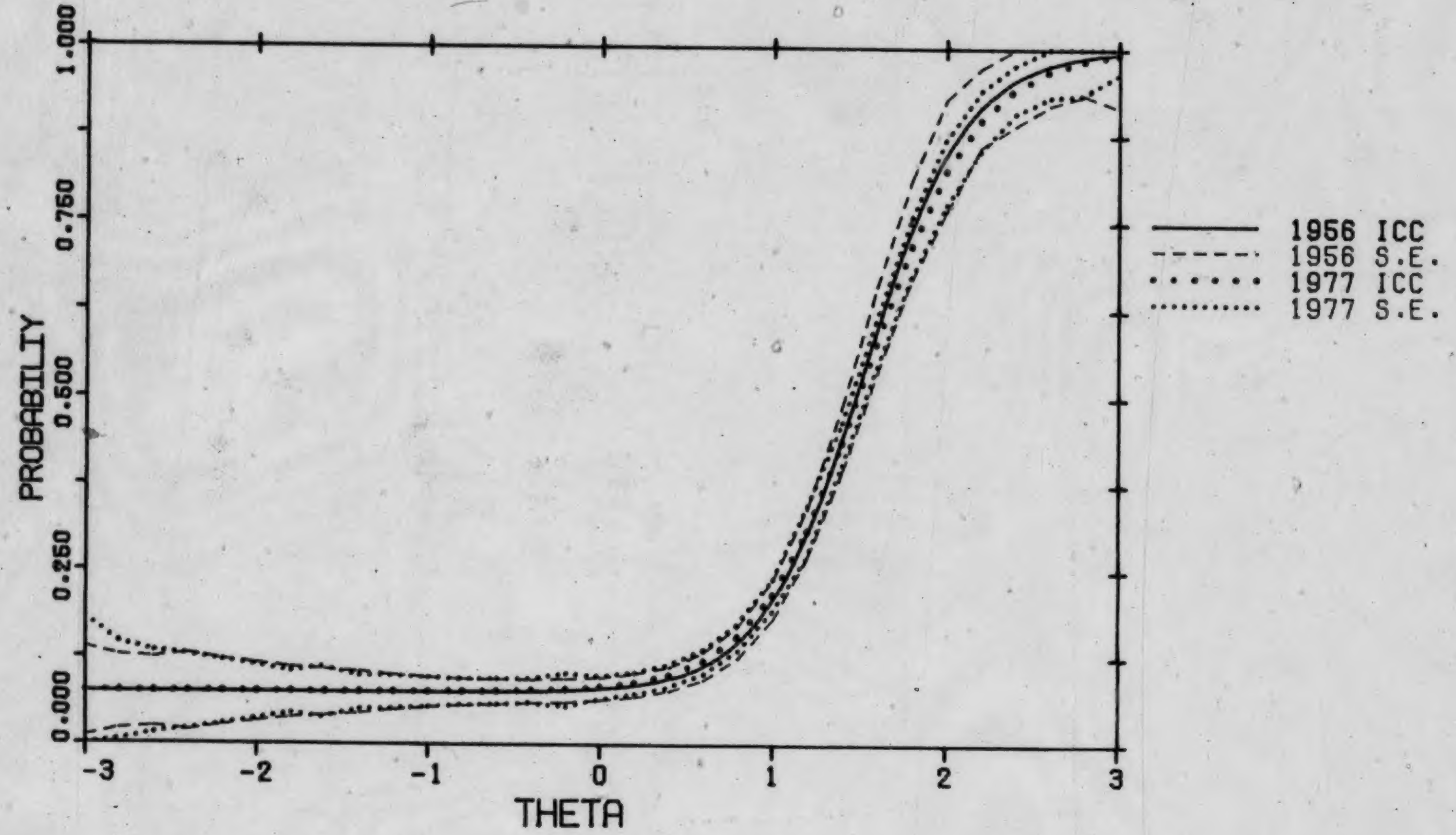


Figure 49

ICC'S FOR ITEM NUMBER 32 : GPR

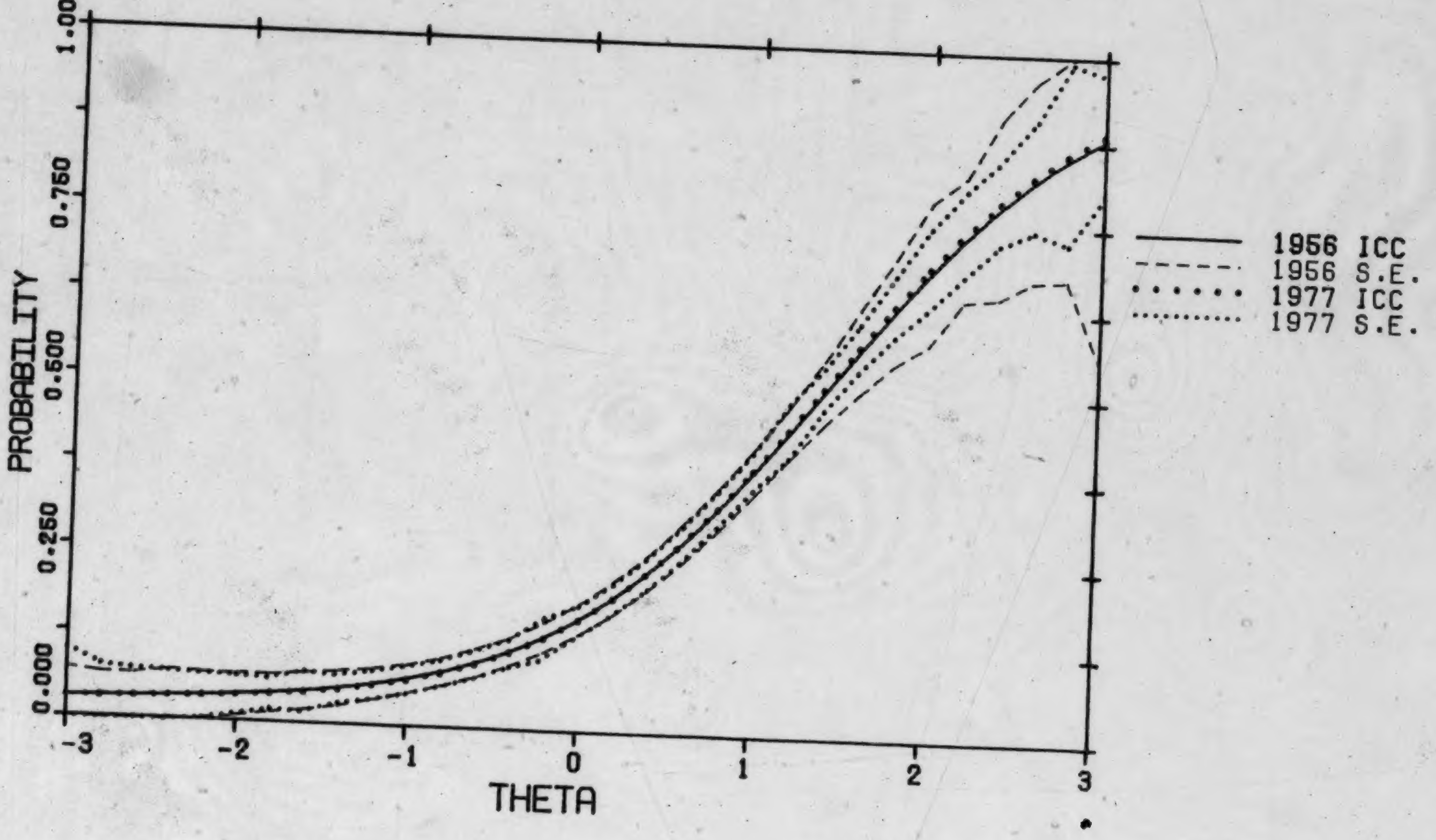


Figure 50

ICC'S FOR ITEM NUMBER 6 : GWR

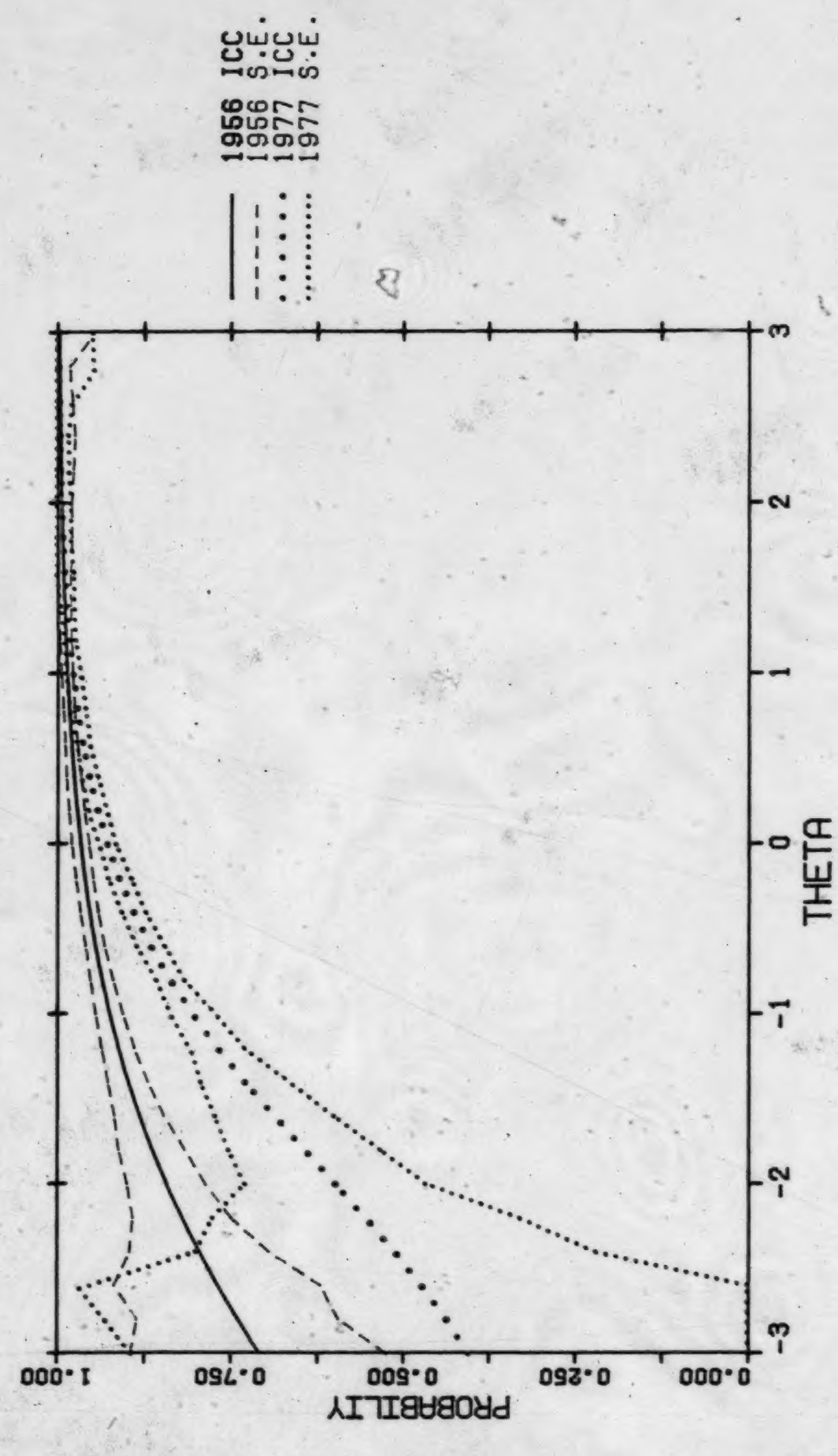


Figure 51

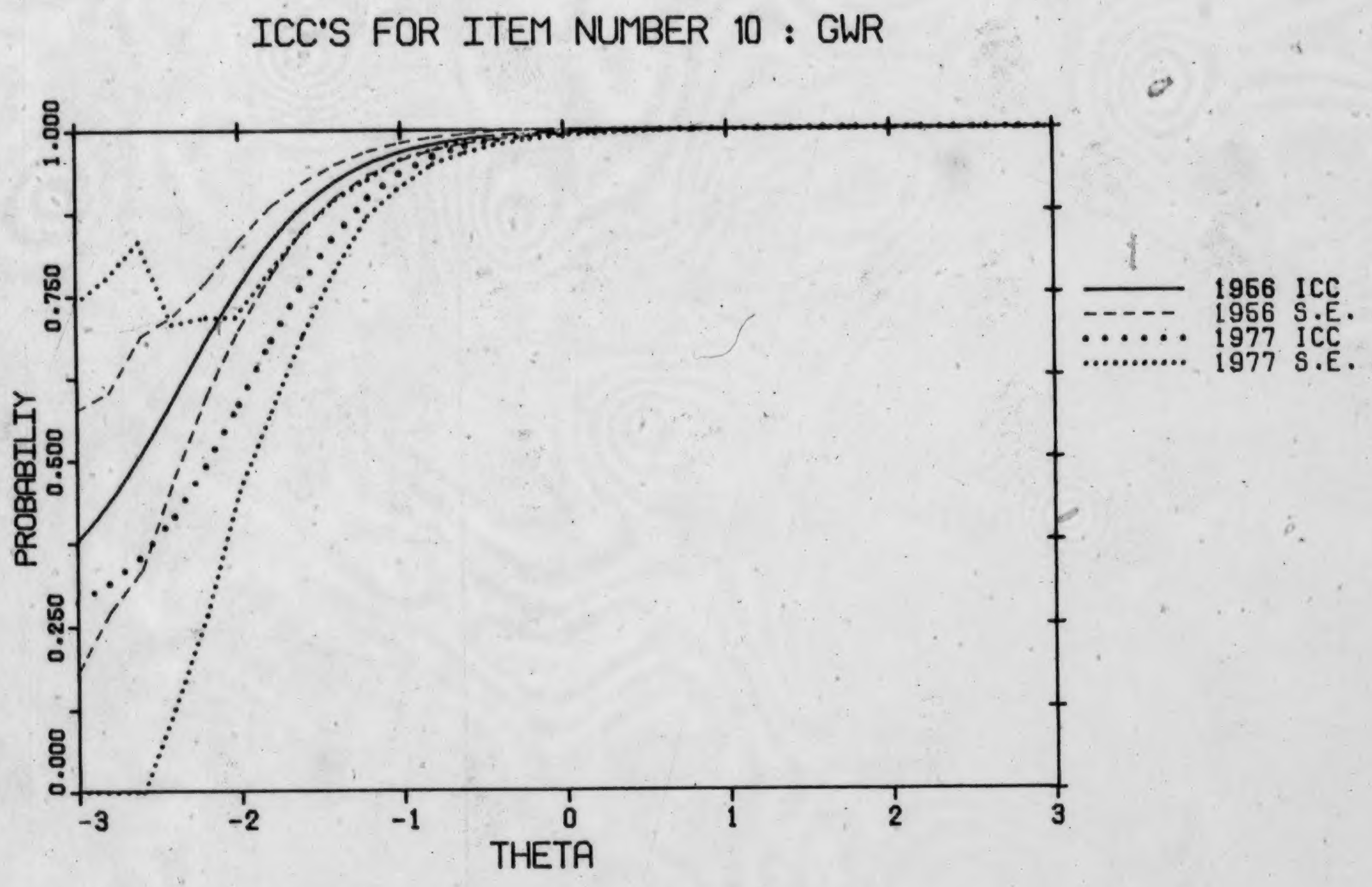

ICC'S FOR ITEM NUMBER 10 : GWR
PROBABILITY
1.000
0.750
0.500
0.250
0.000
-3
-2
-1
0
1
2
3
THETA
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 52

ICC'S FOR ITEM NUMBER 11 : GWR

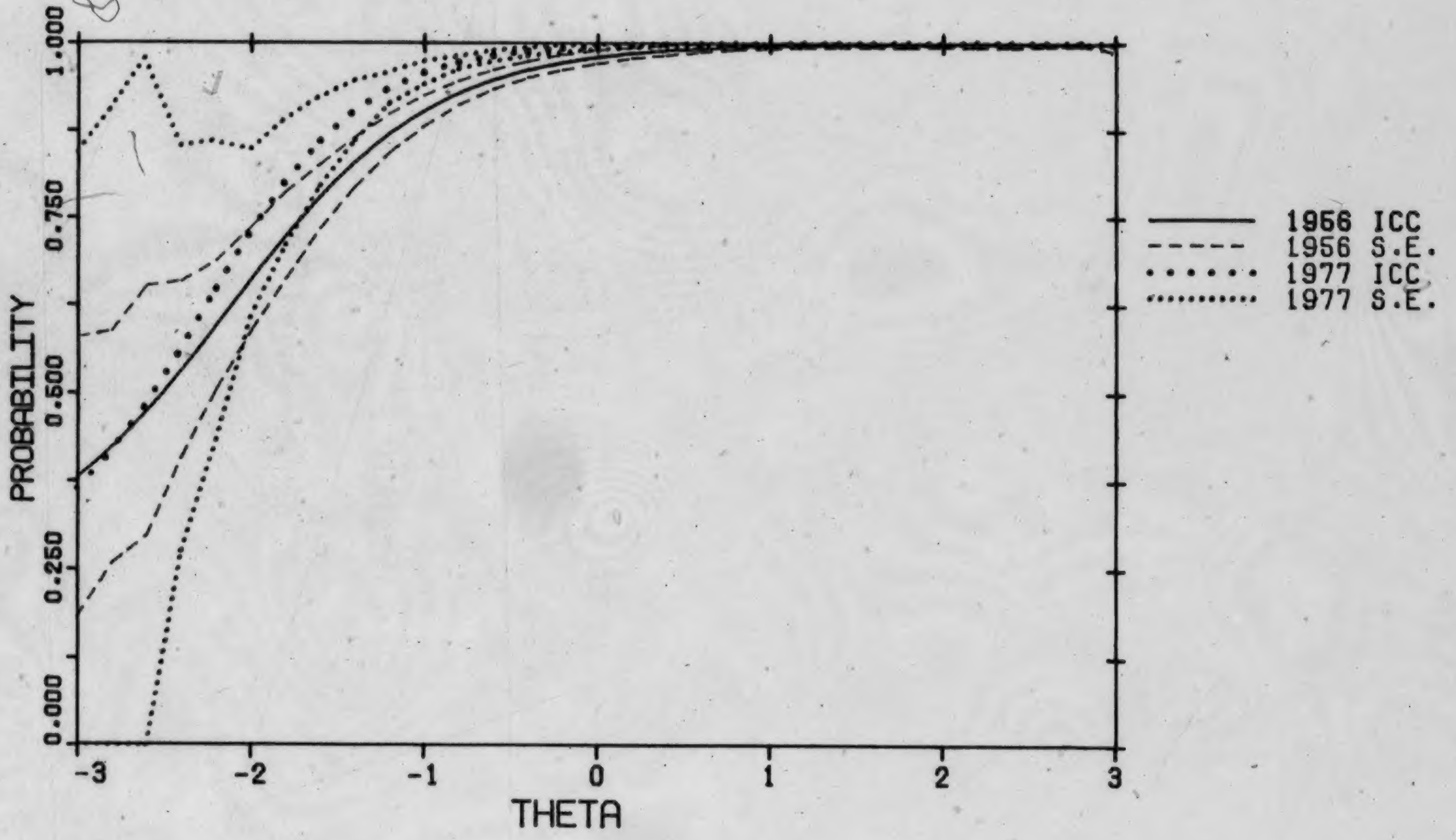


Figure 53

ICC'S FOR ITEM NUMBER 12 : GWR

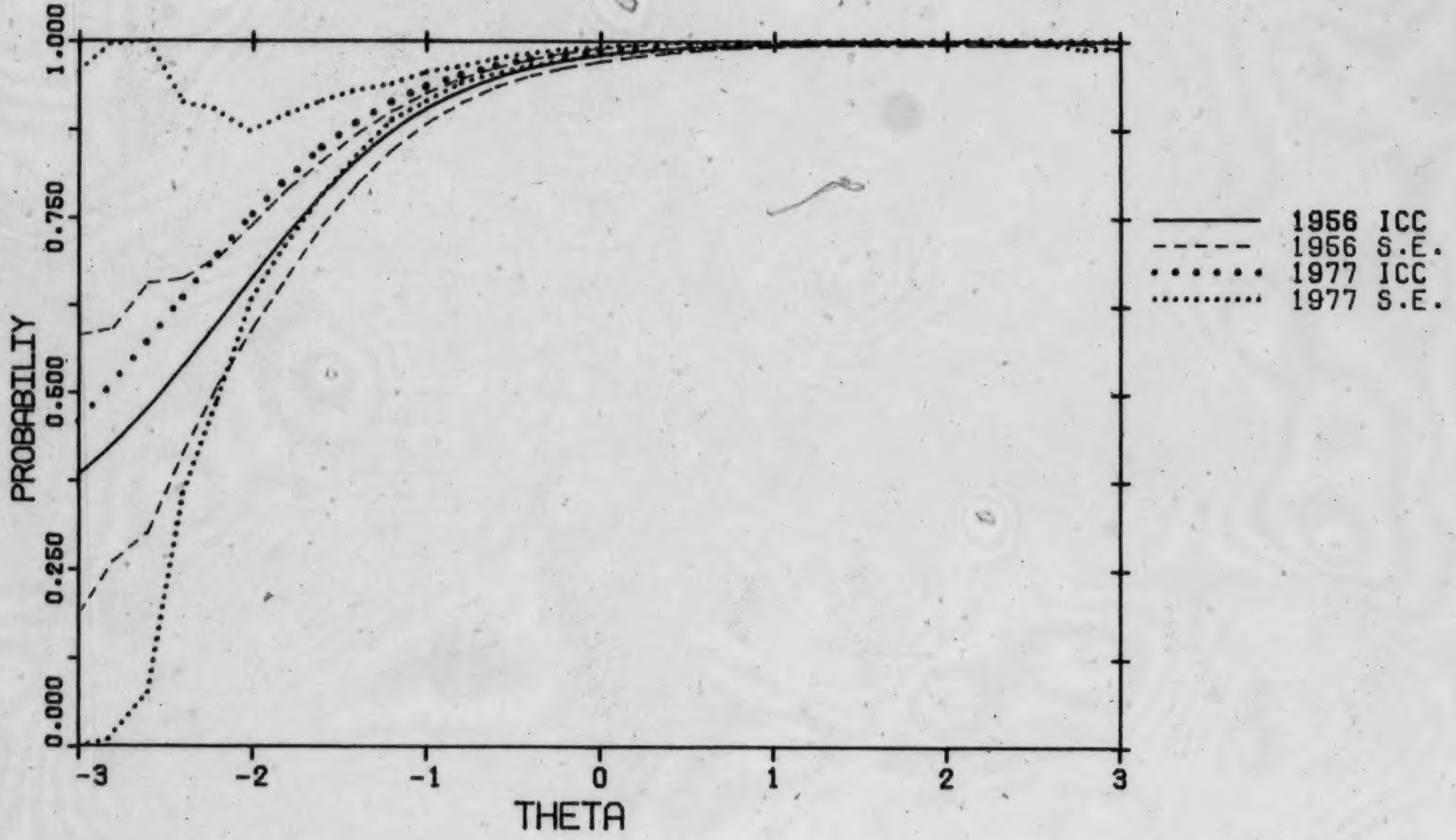


Figure 54

ICC'S FOR ITEM NUMBER 13 : GWR

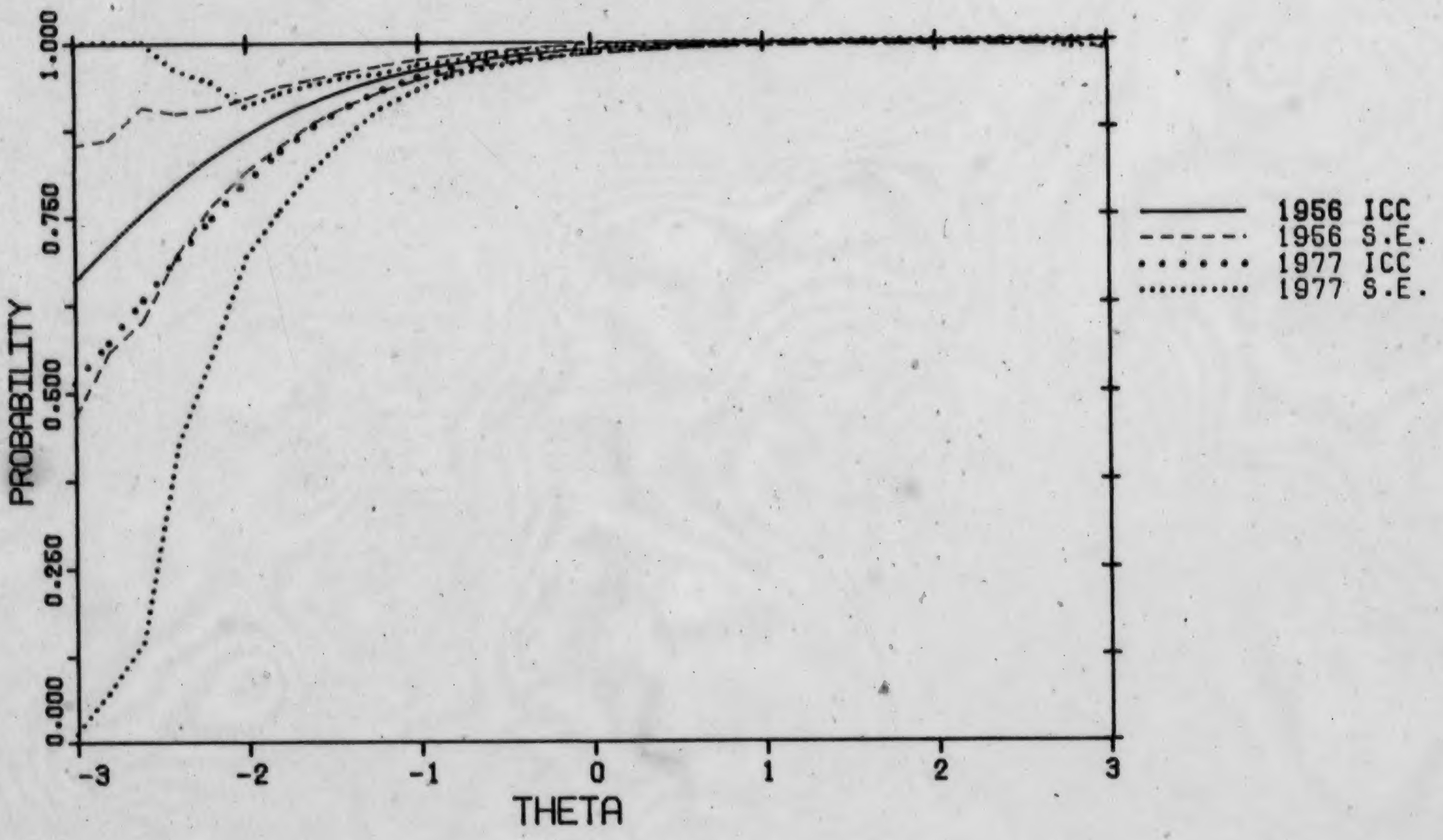


Figure 55

ICC'S FOR ITEM NUMBER 14 : GWR

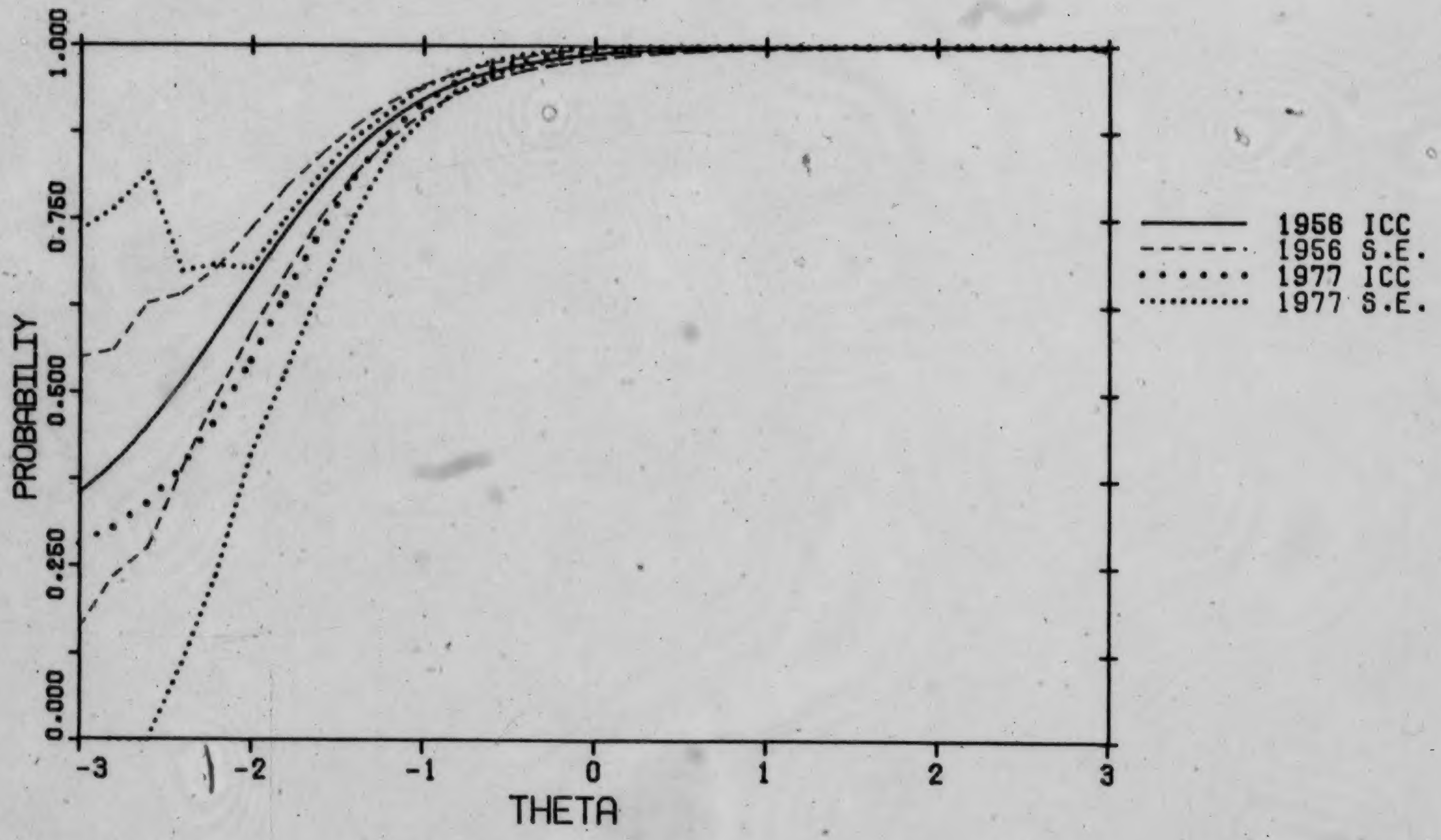


Figure 56

ICC'S FOR ITEM NUMBER 15 : GWR

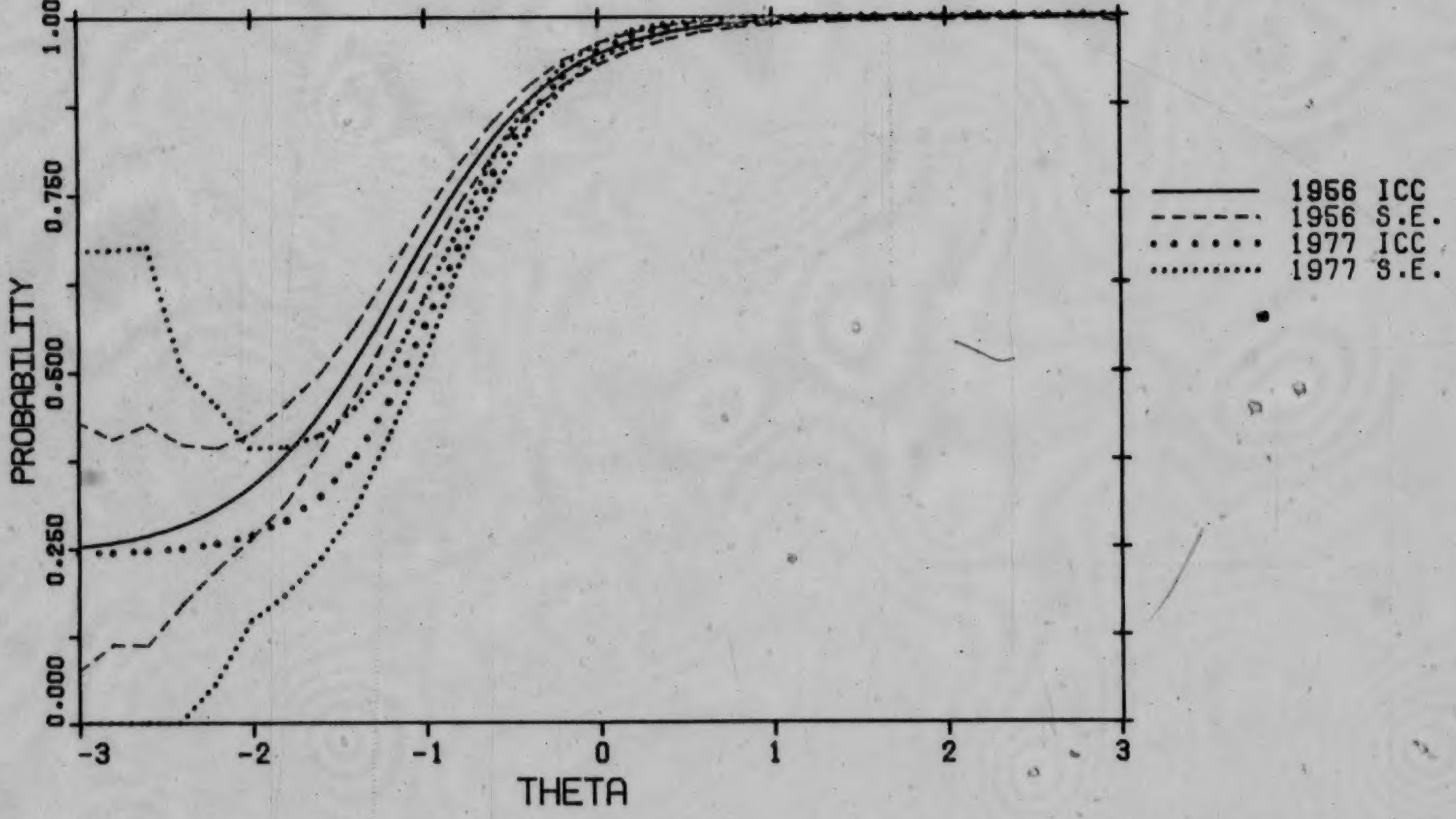


Figure 57

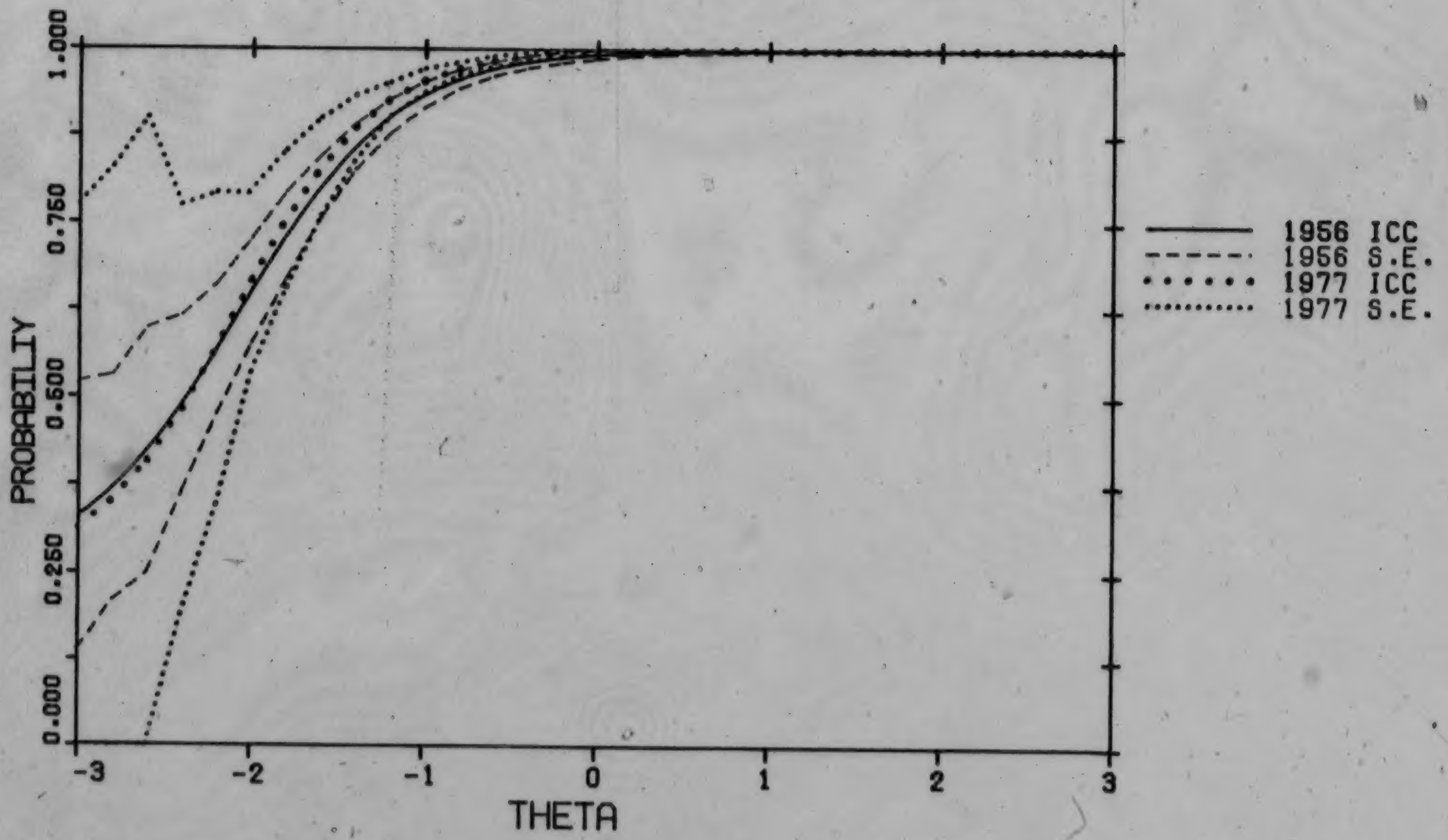

ICC'S FOR ITEM NUMBER 16 : GWR

Figure 58

ICC'S FOR ITEM NUMBER 17 : GWR

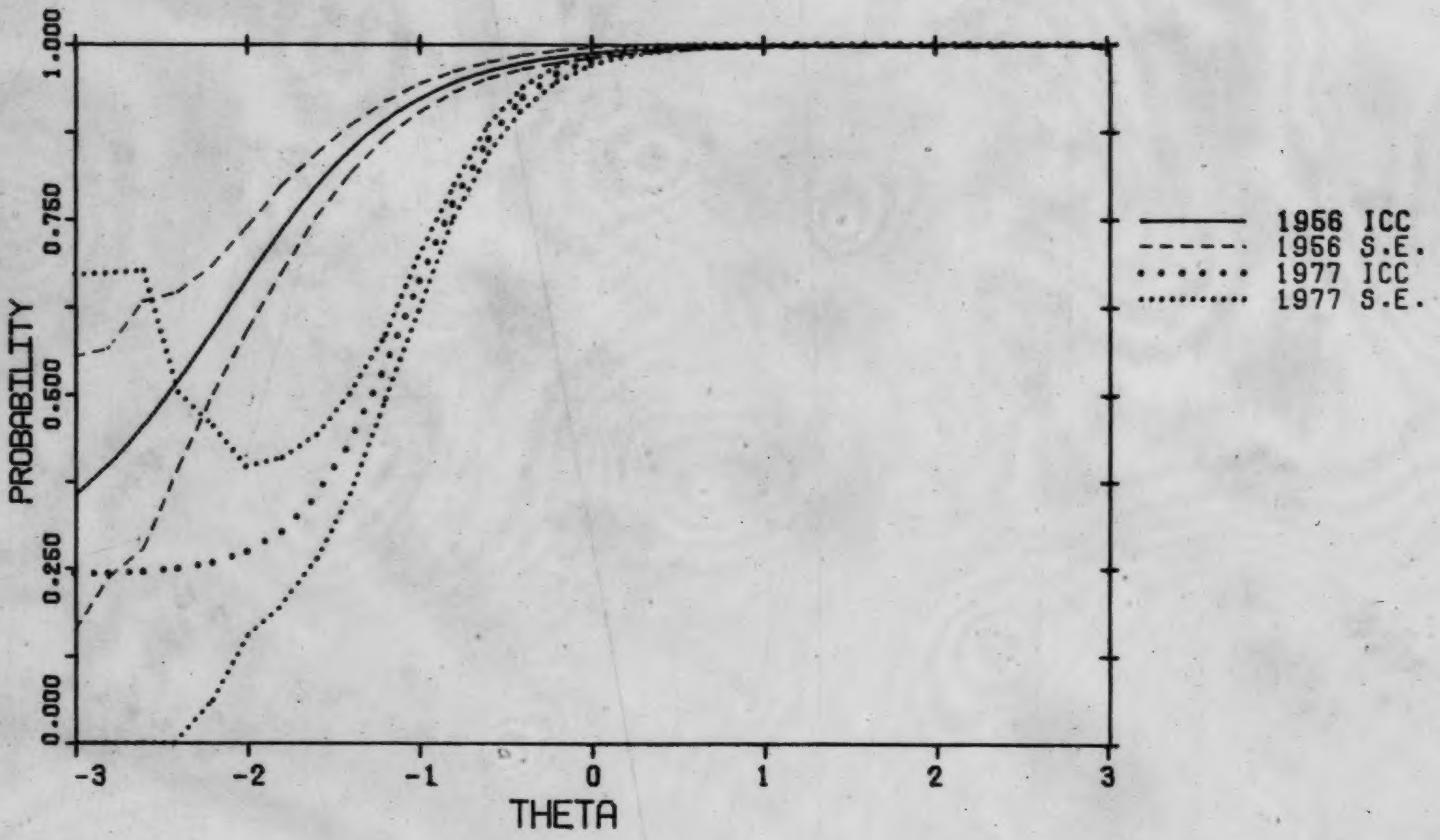


Figure 59

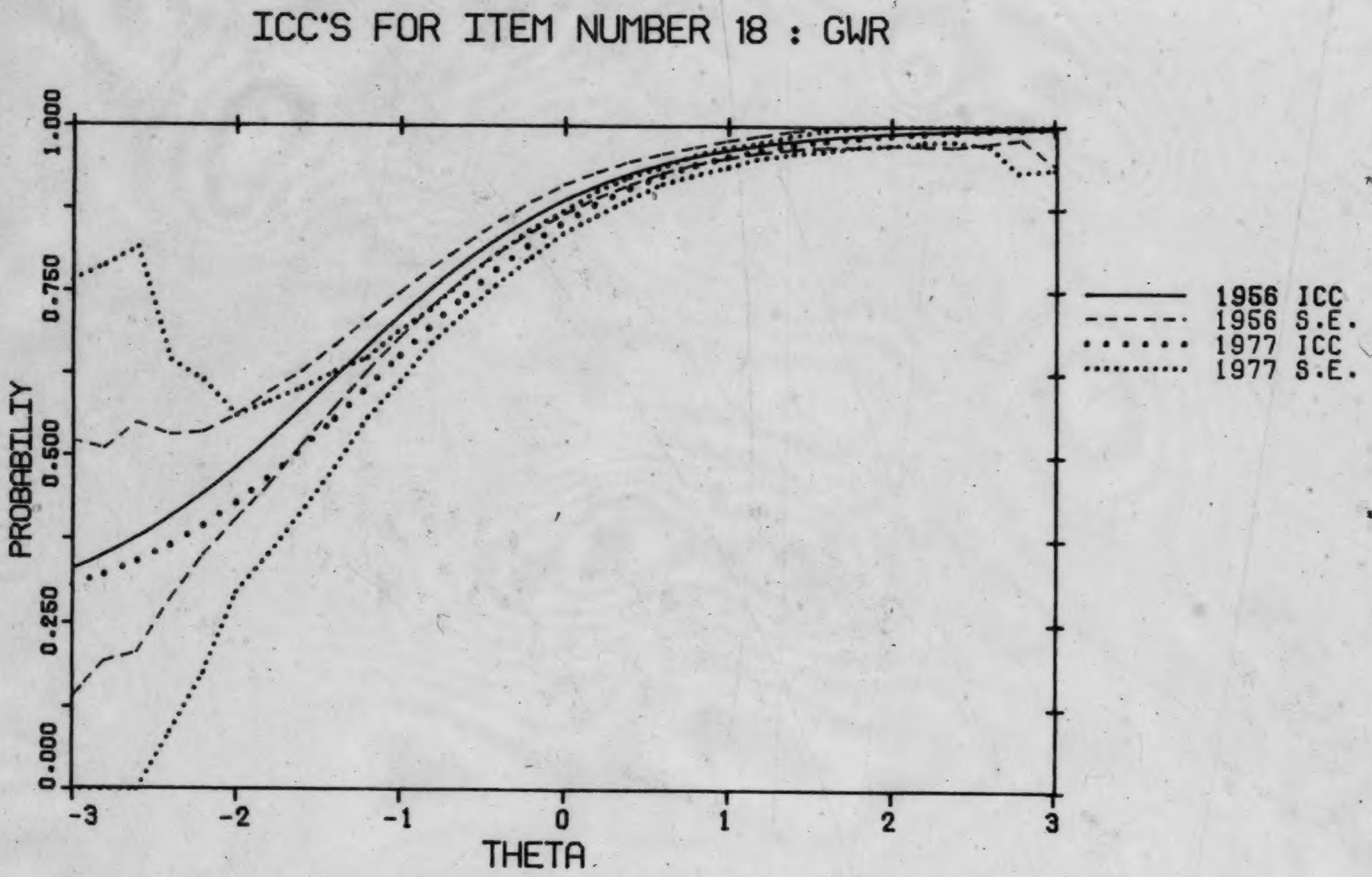
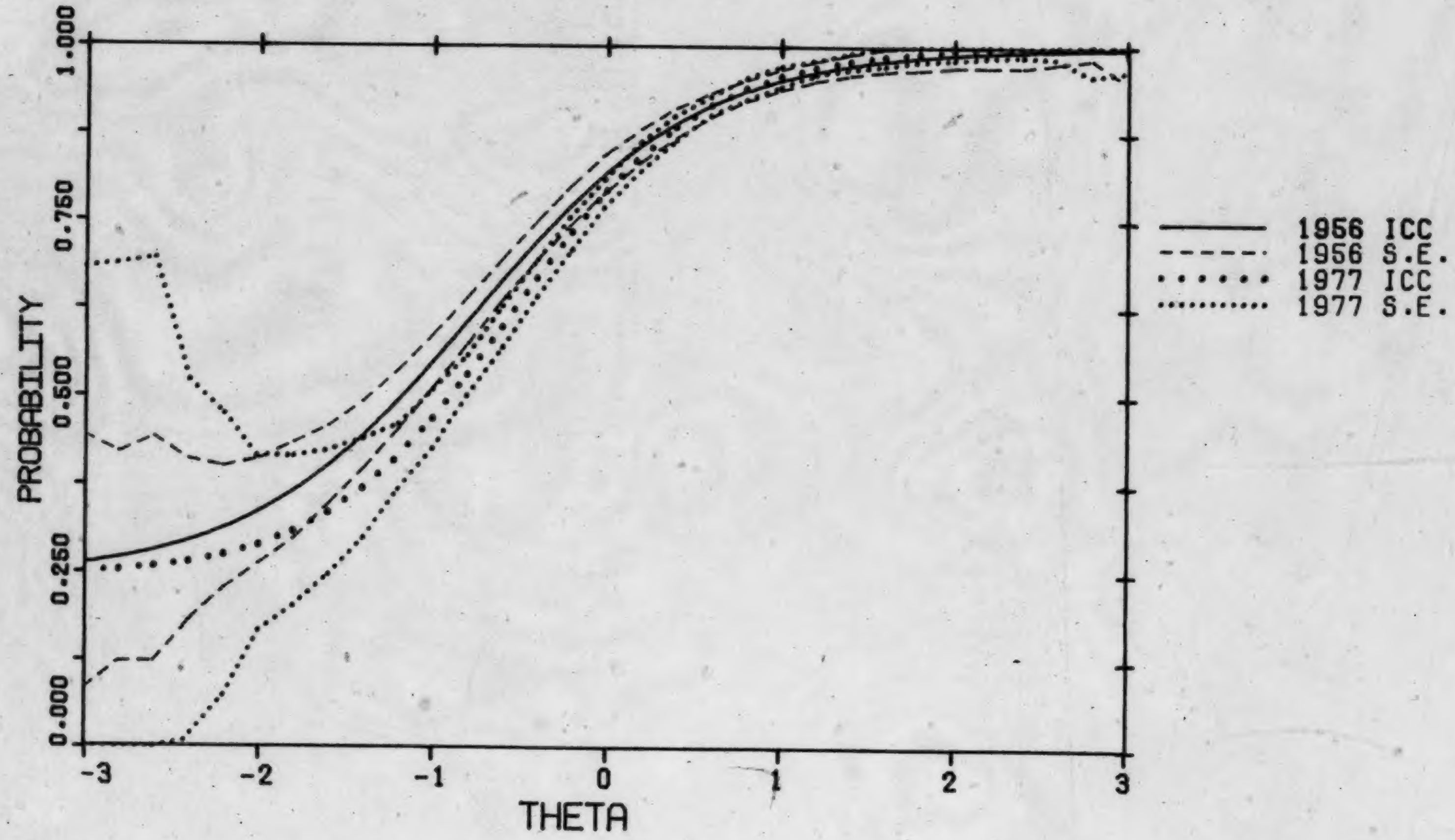

ICC'S FOR ITEM NUMBER 18 : GWR
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 60



ICC'S FOR ITEM NUMBER 19 : GWR

Figure 61

ICC'S FOR ITEM NUMBER 20 : GWR

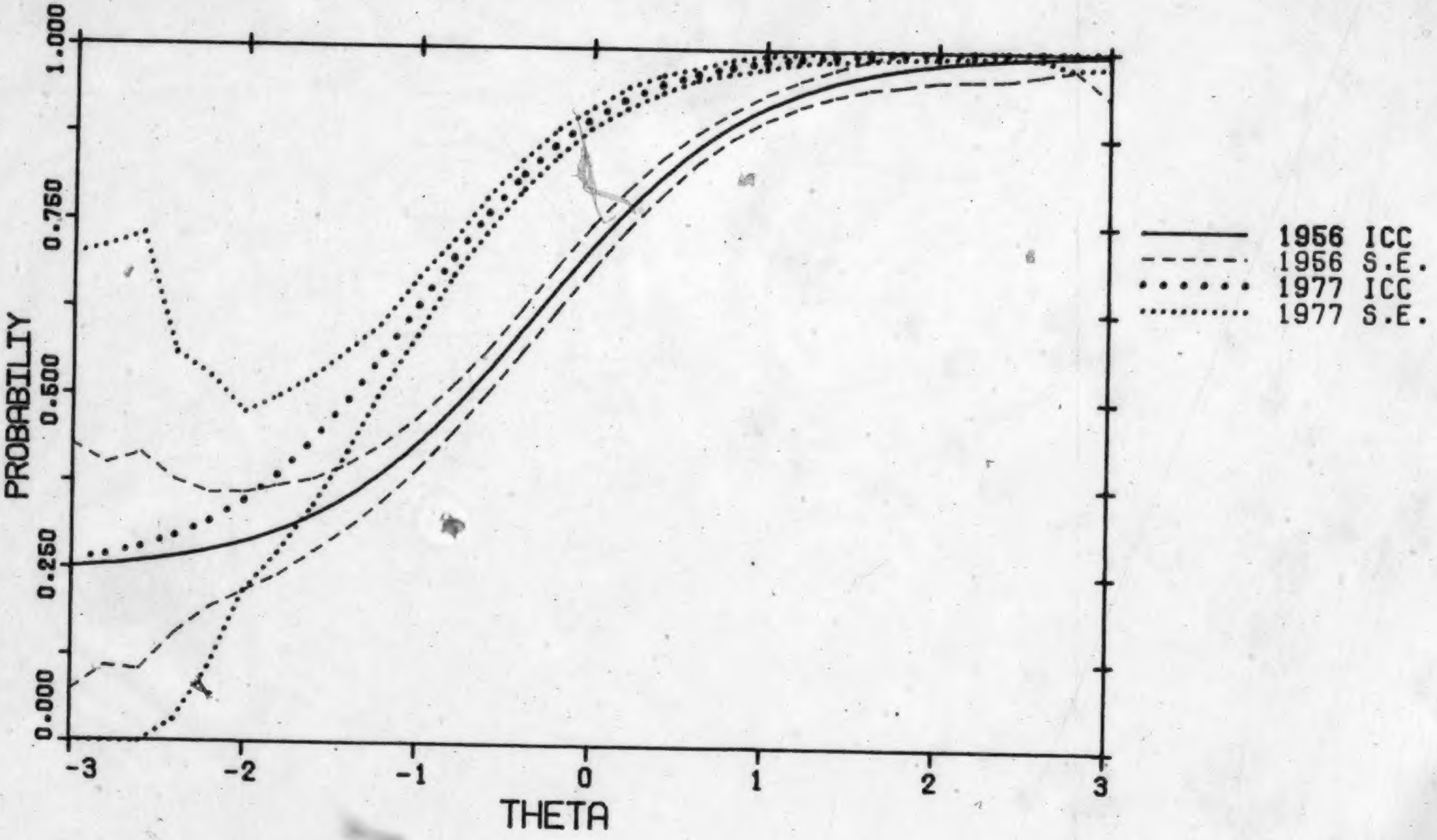


Figure 62

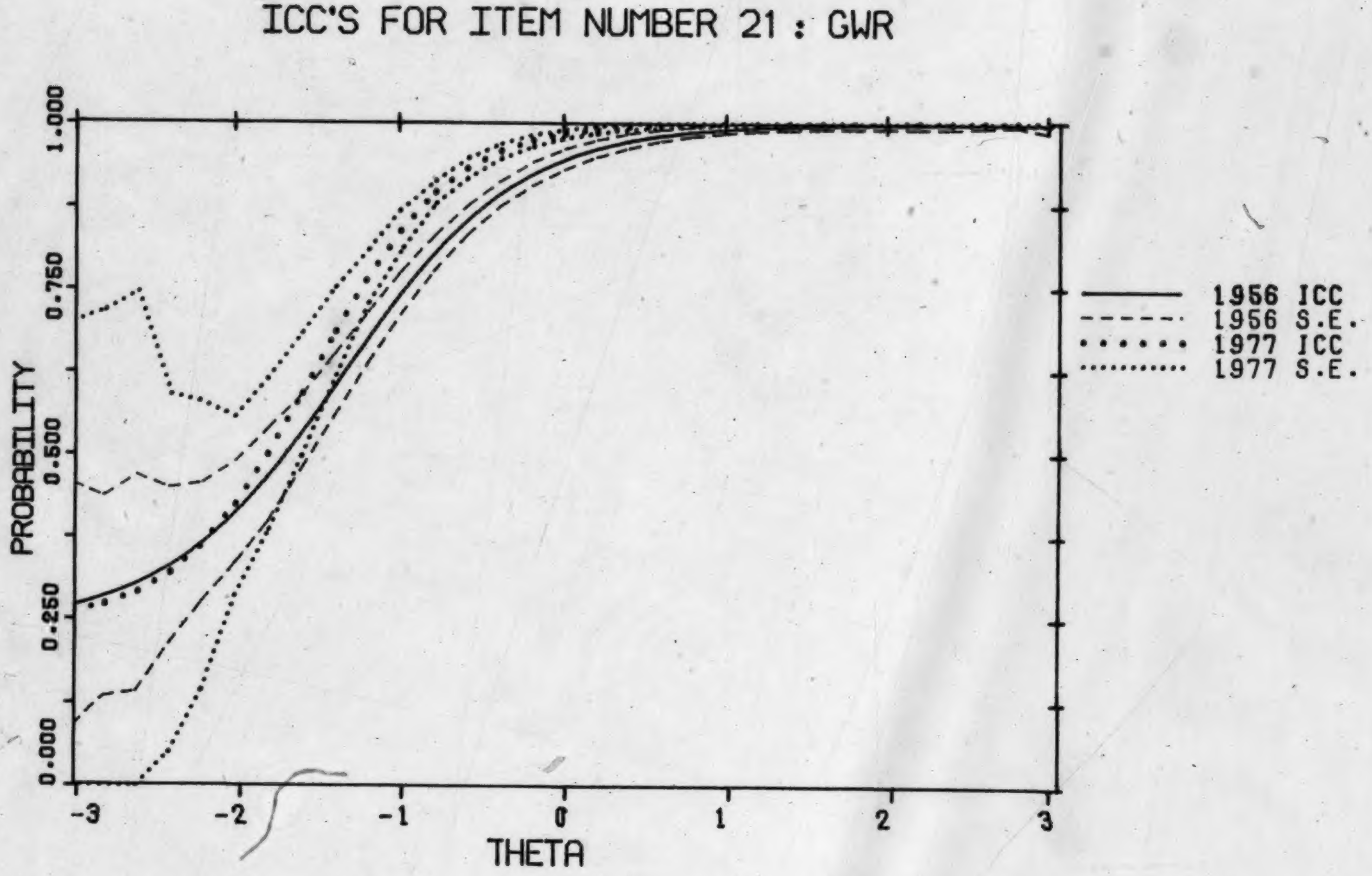

ICC'S FOR ITEM NUMBER 21 : GWR
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 63

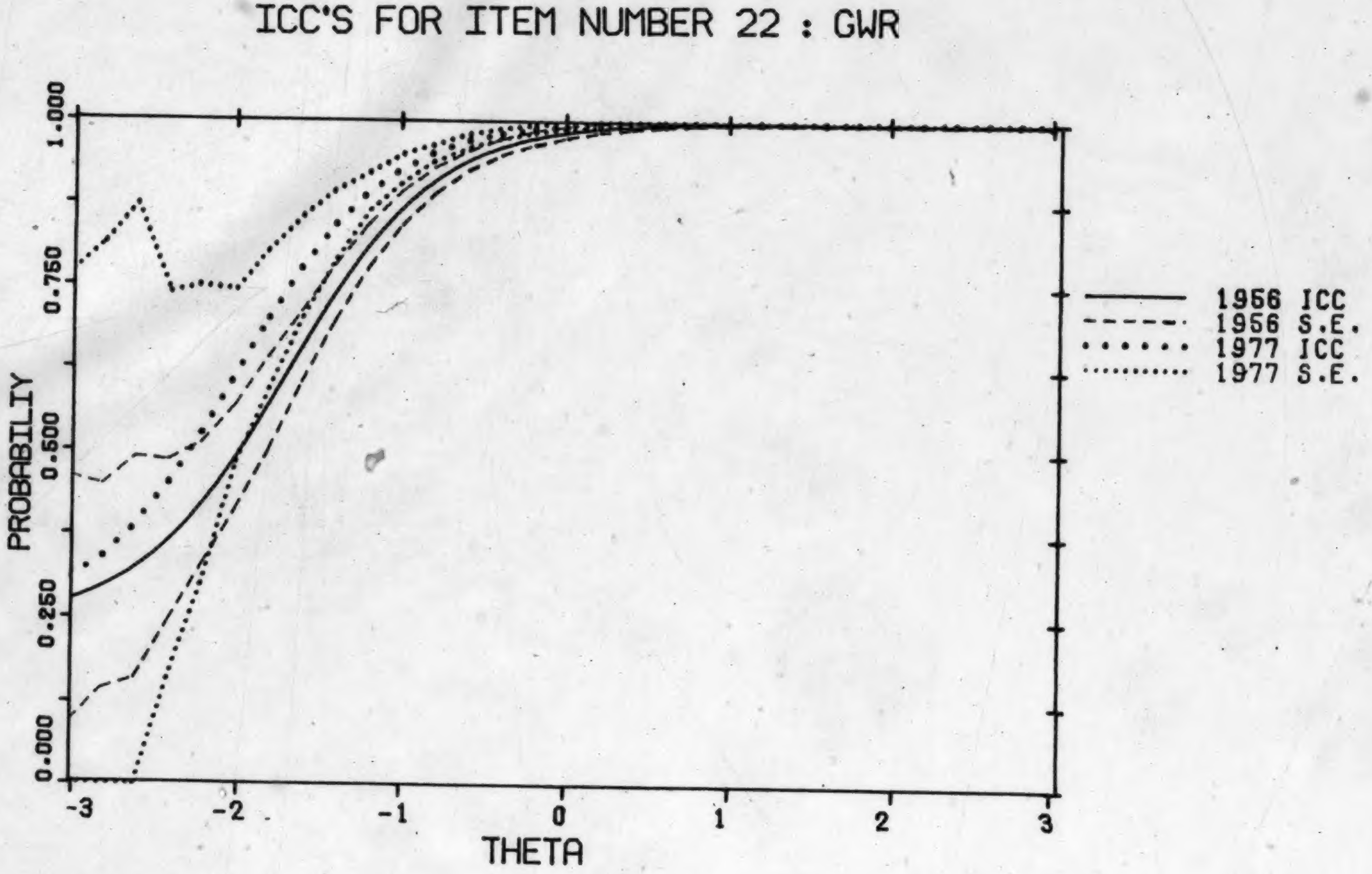

ICC'S FOR ITEM NUMBER 22 : GWR
PROBABILIY
1.000
0.750
0.500
0.250
0.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 64

ICC'S FOR ITEM NUMBER 23 : GWR

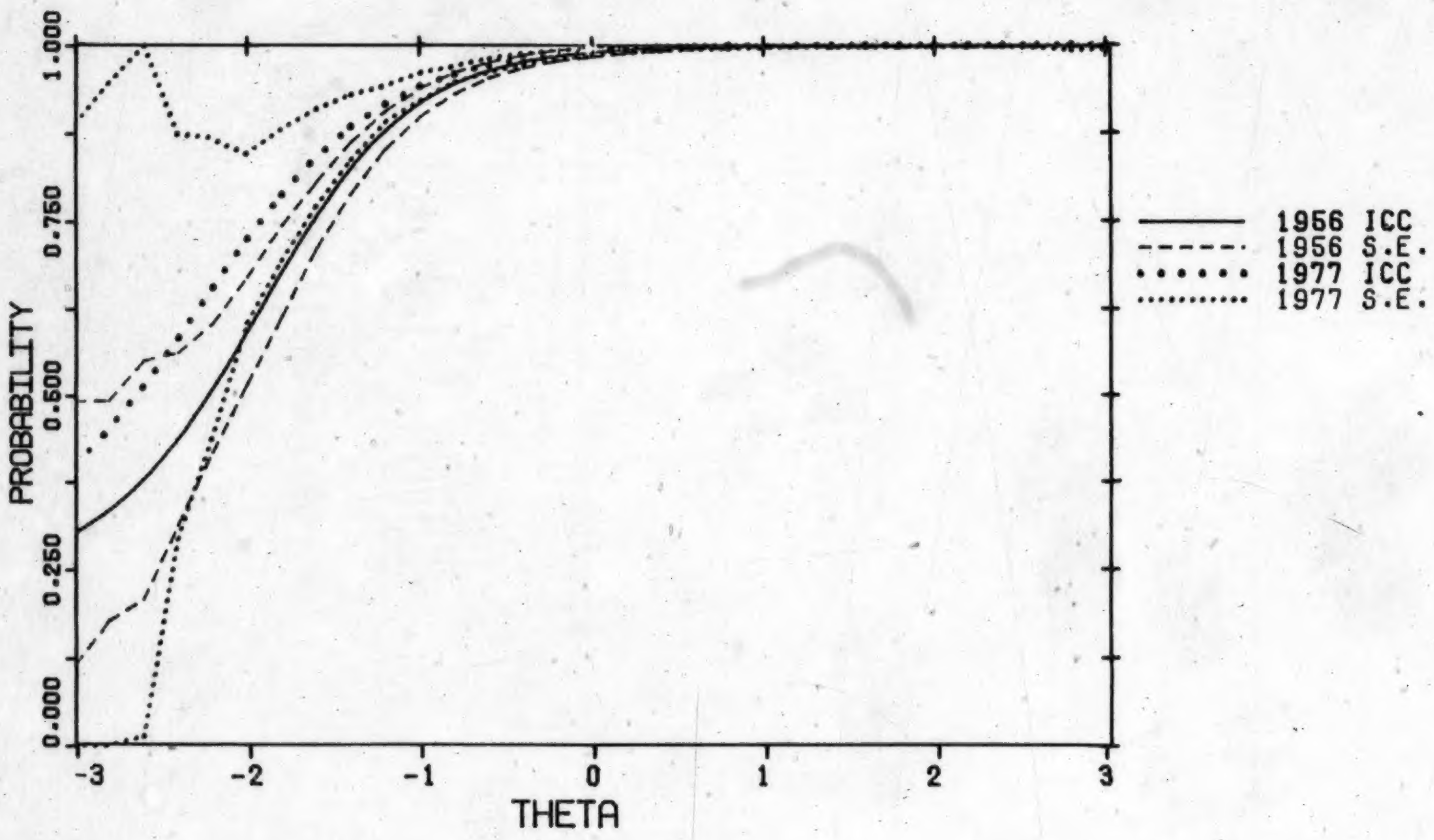


Figure 65

ICC'S FOR ITEM NUMBER 24 : GWR

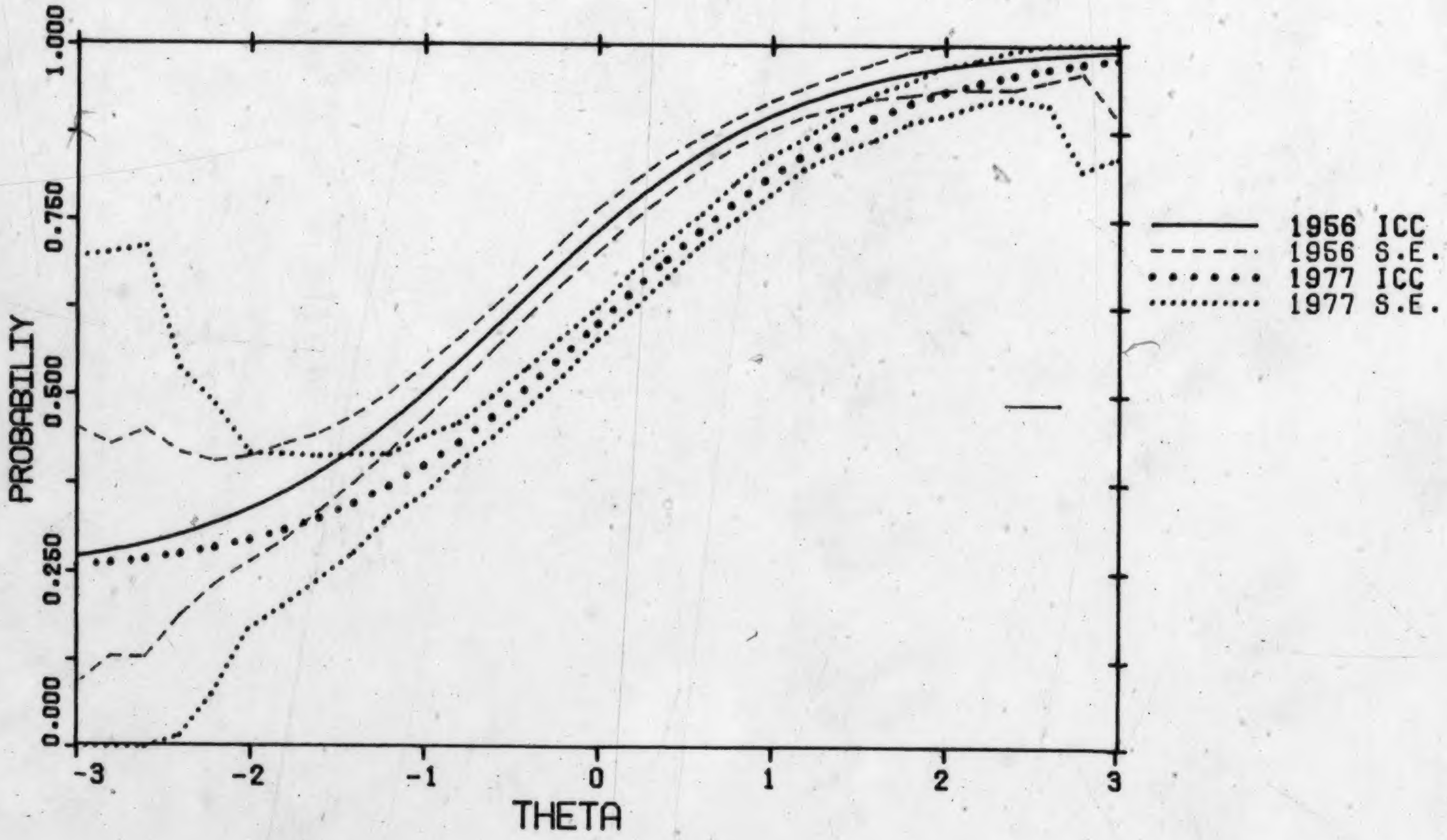


Figure 66

ICC'S FOR ITEM NUMBER 25 : GWR

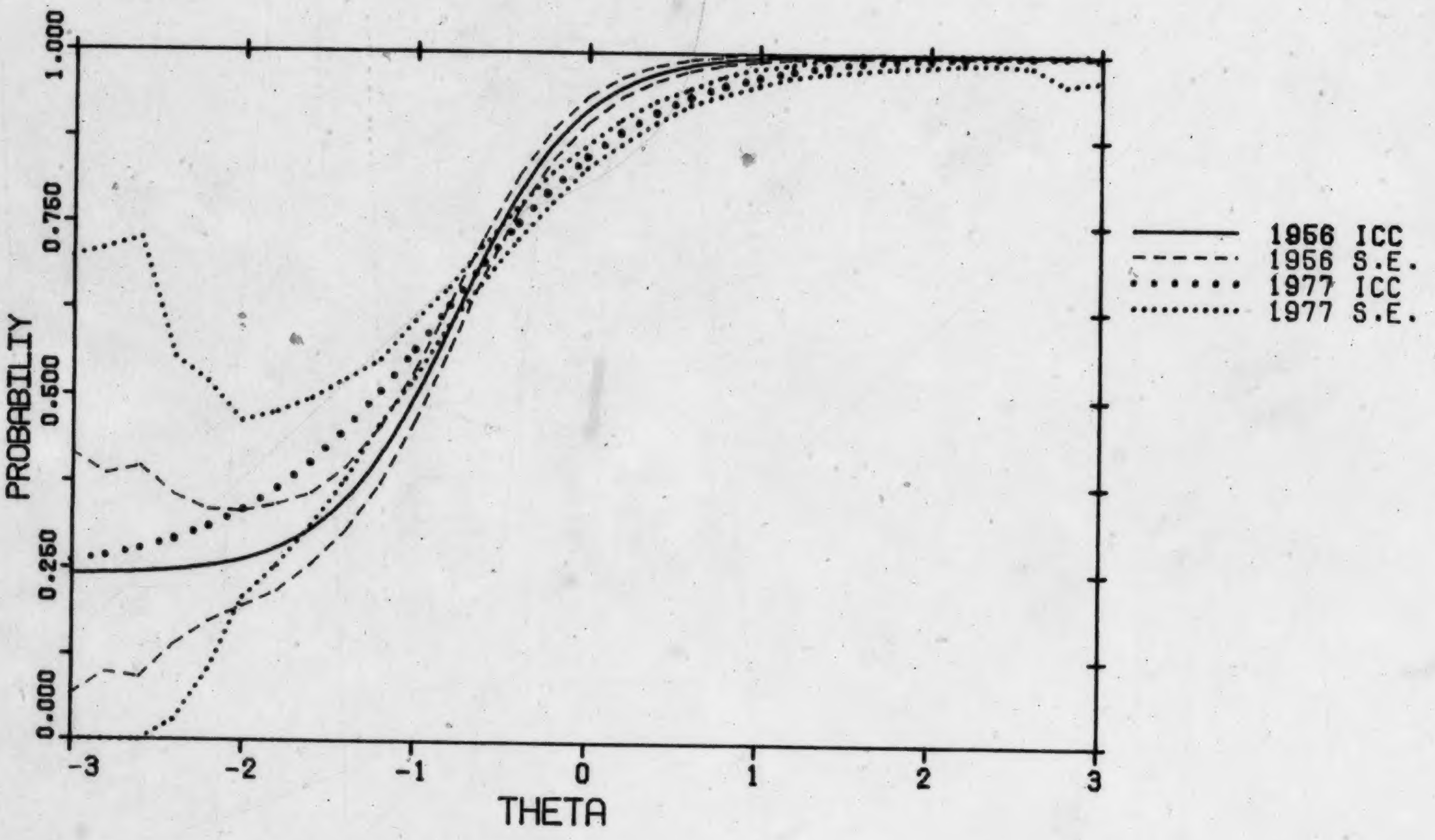


Figure 67

ICC'S FOR ITEM NUMBER 26 : GWR

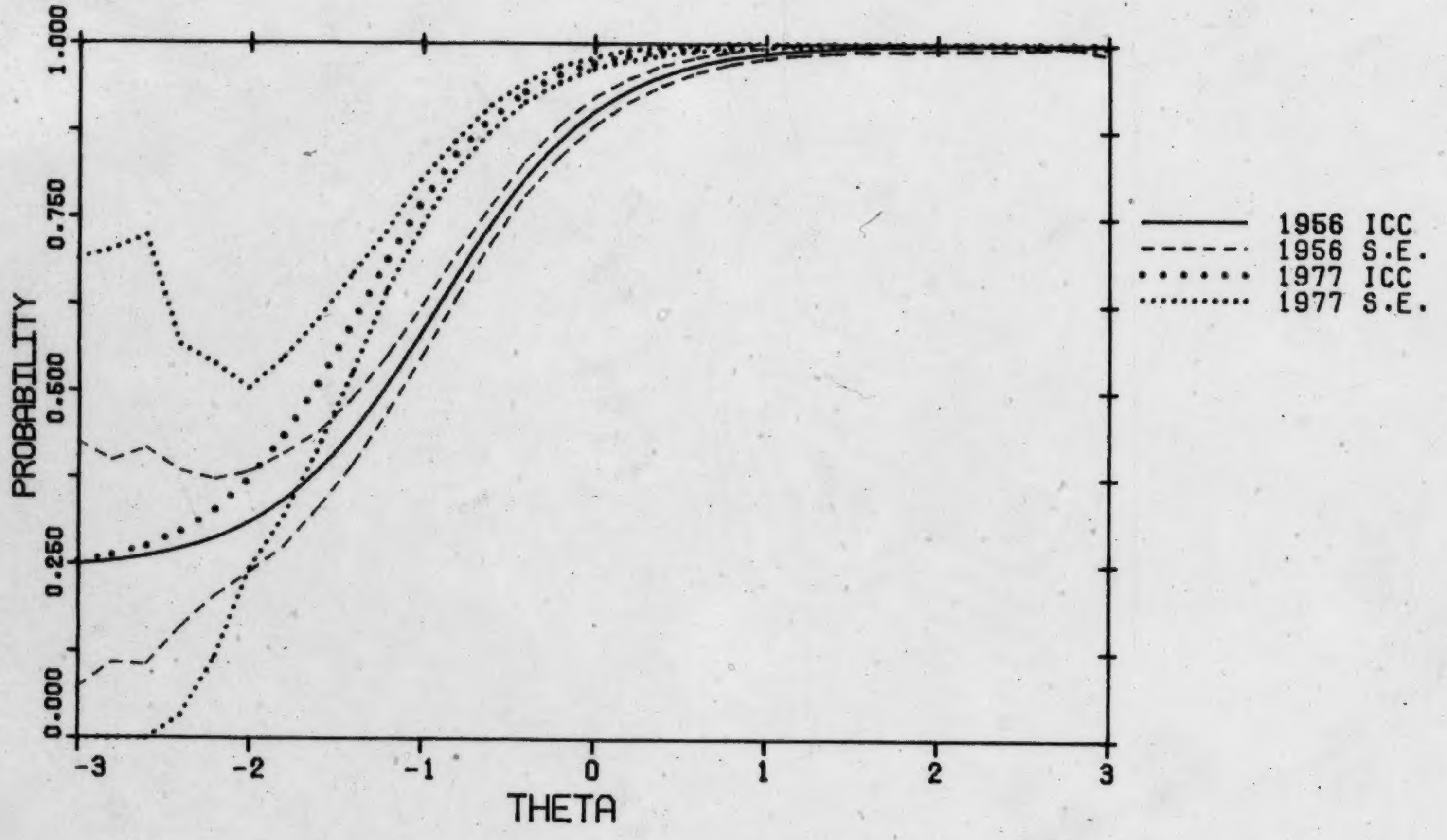


Figure 68

ICC'S FOR ITEM NUMBER 27 : GWR

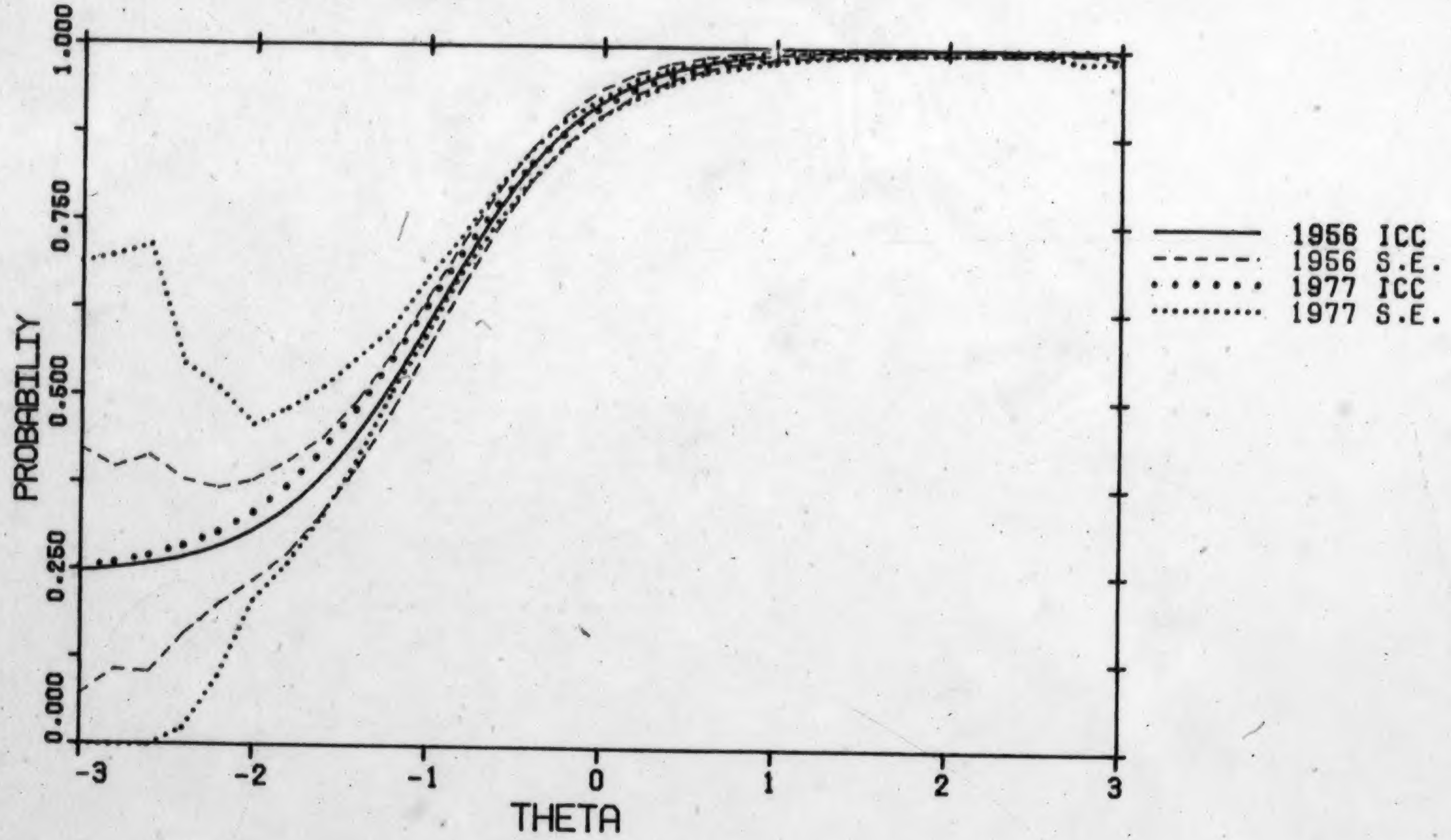


Figure 69

ICC'S FOR ITEM NUMBER 28 : GWR

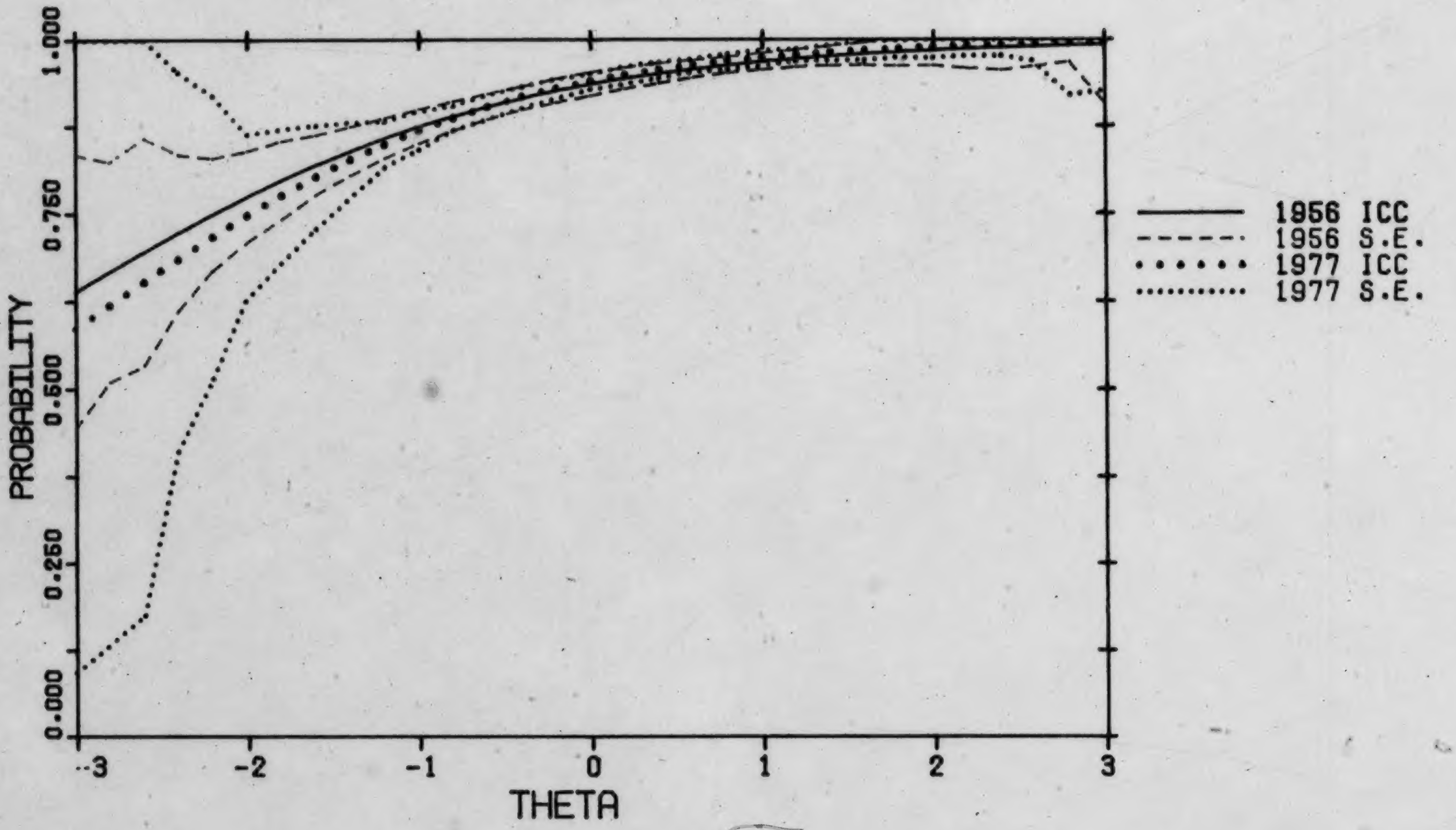


Figure 70

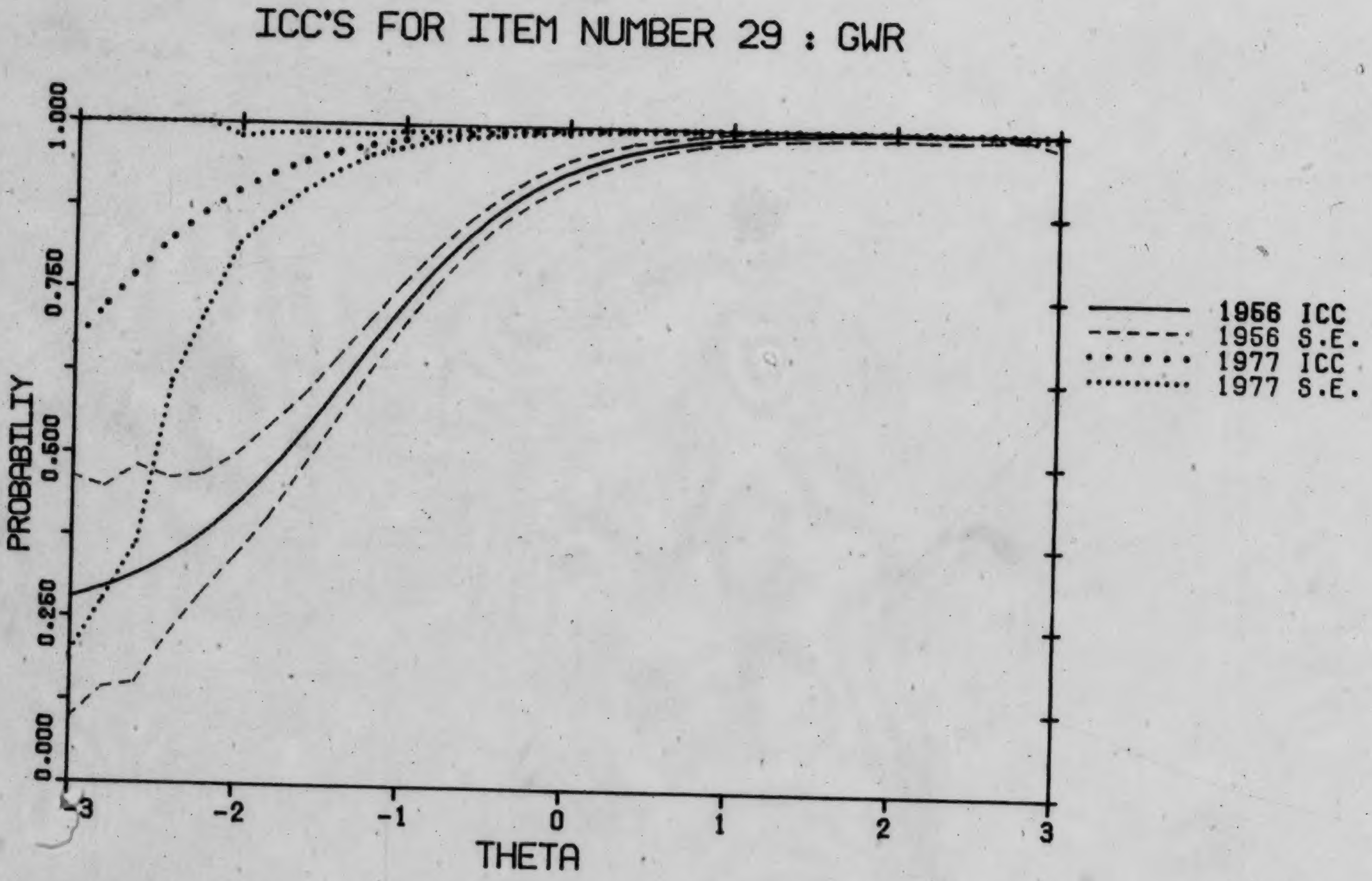

ICC'S FOR ITEM NUMBER 29 : GWR
PROBABILIY
1.000
0.750
0.500
0.250
0.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 71

ICC'S FOR ITEM NUMBER 30 : GWR

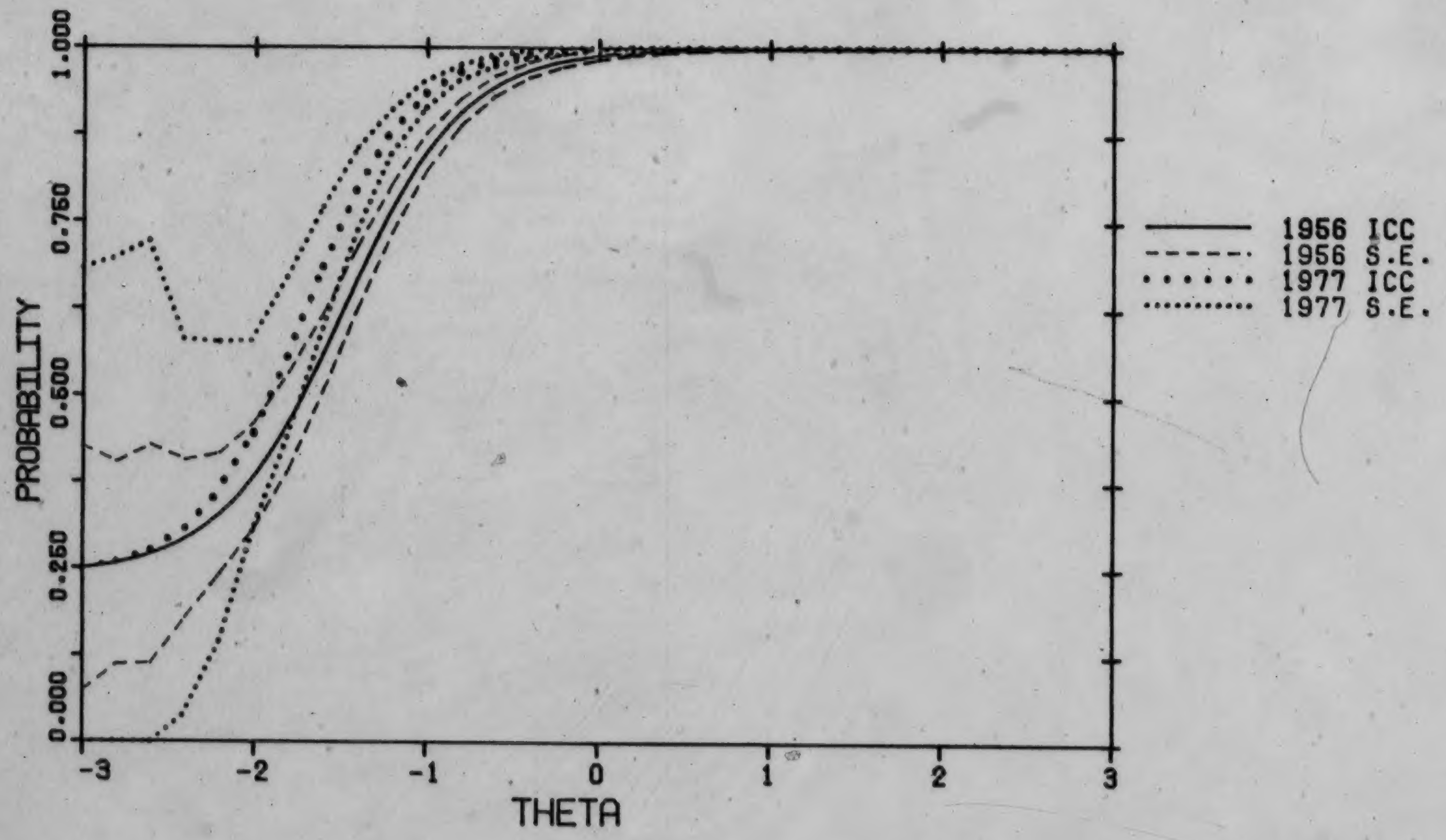


Figure 72

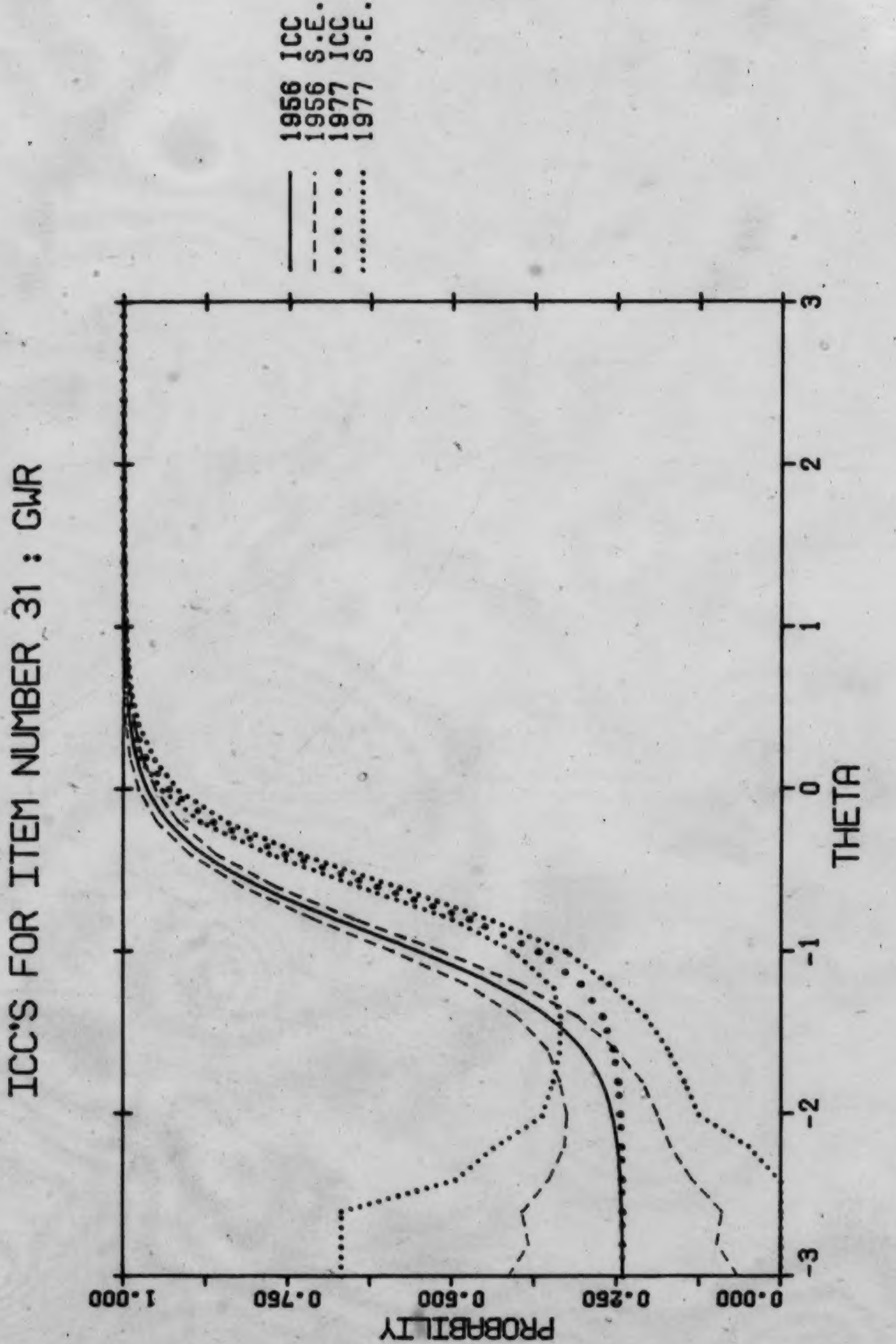

ICC'S FOR ITEM NUMBER 31 : GWR
PROBABILITY
1.000
0.750
0.500
0.250
0.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 73

ICC'S FOR ITEM NUMBER 32 : GWR

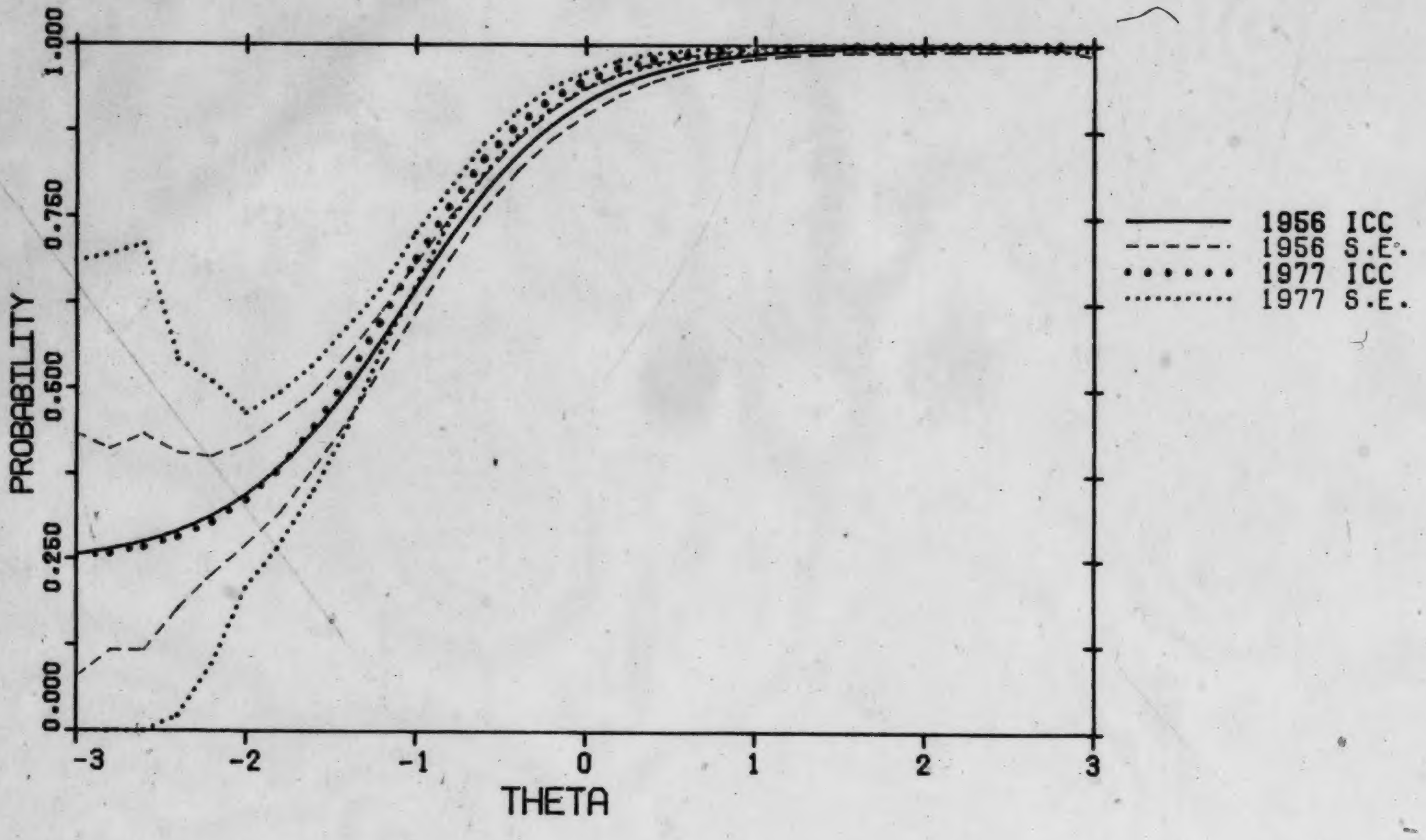


Figure 74

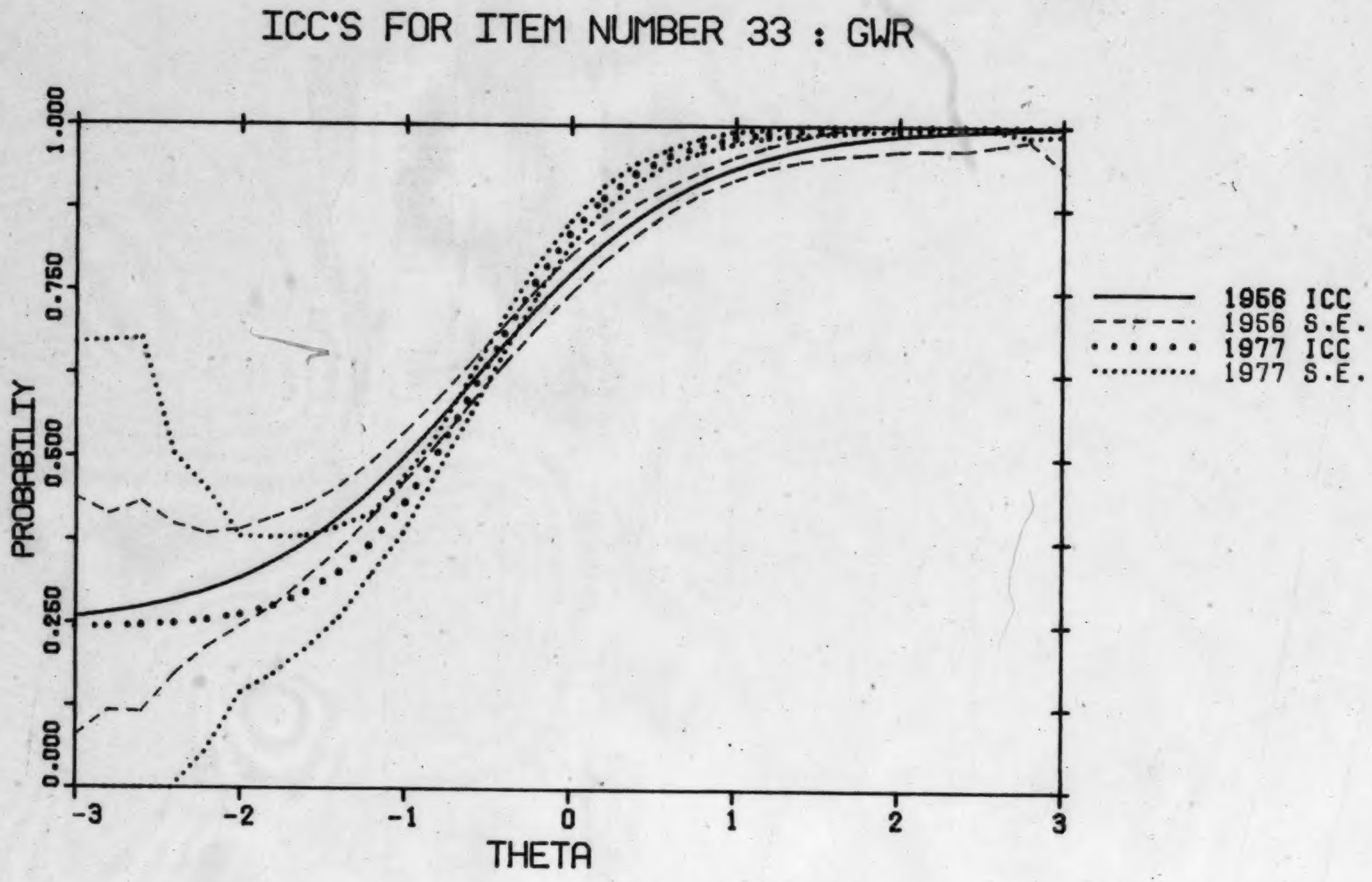
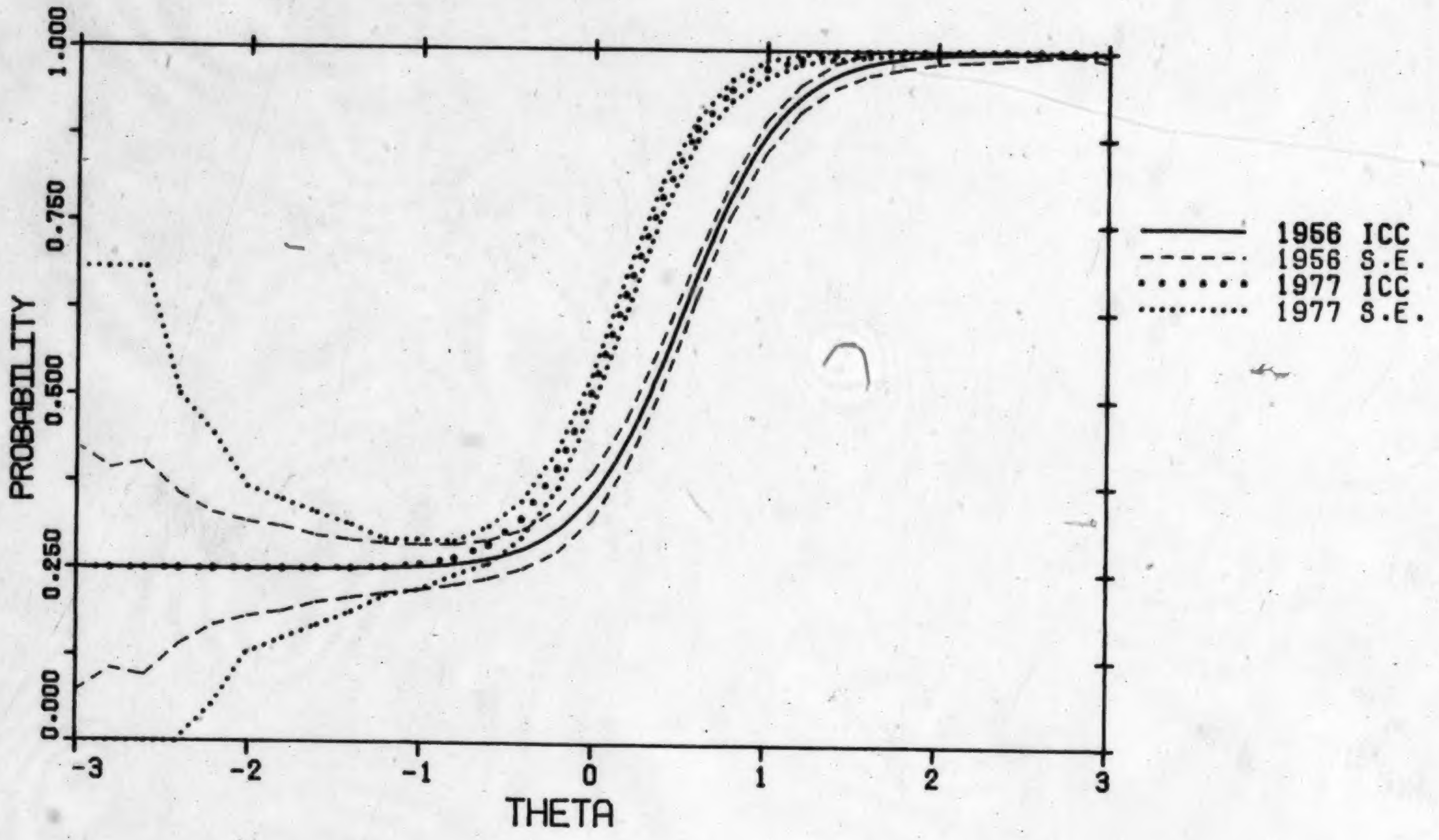

ICC'S FOR ITEM NUMBER 33 : GWR
PROBABILITY
0.000
0.250
0.500
0.750
1.000
-3
-2
-1
0
1
2
3
THETA
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 75



ICC'S FOR ITEM NUMBER 34 : GWR

Figure 76

ICC'S FOR ITEM NUMBER 35 : GWR

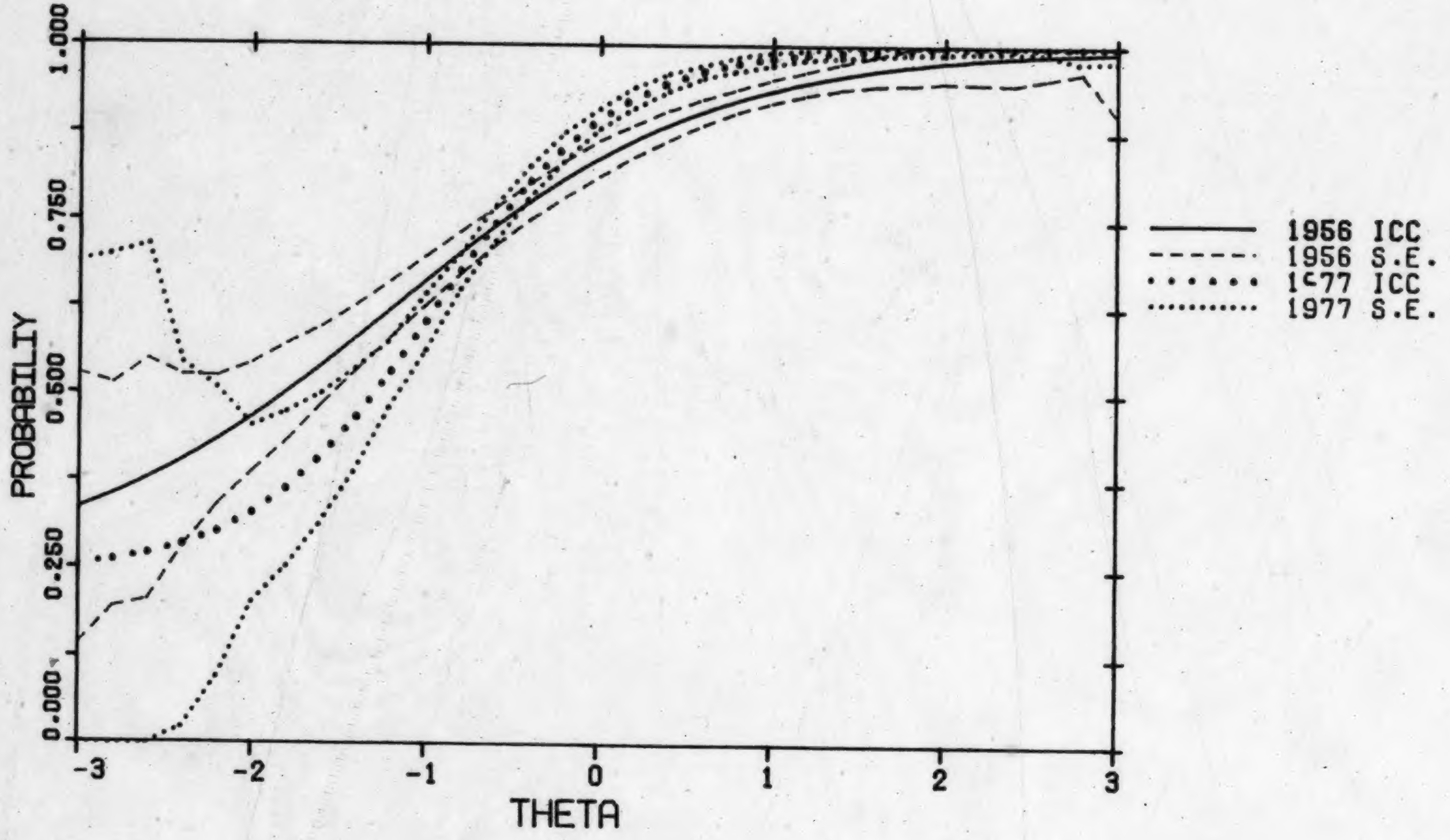


Figure 77

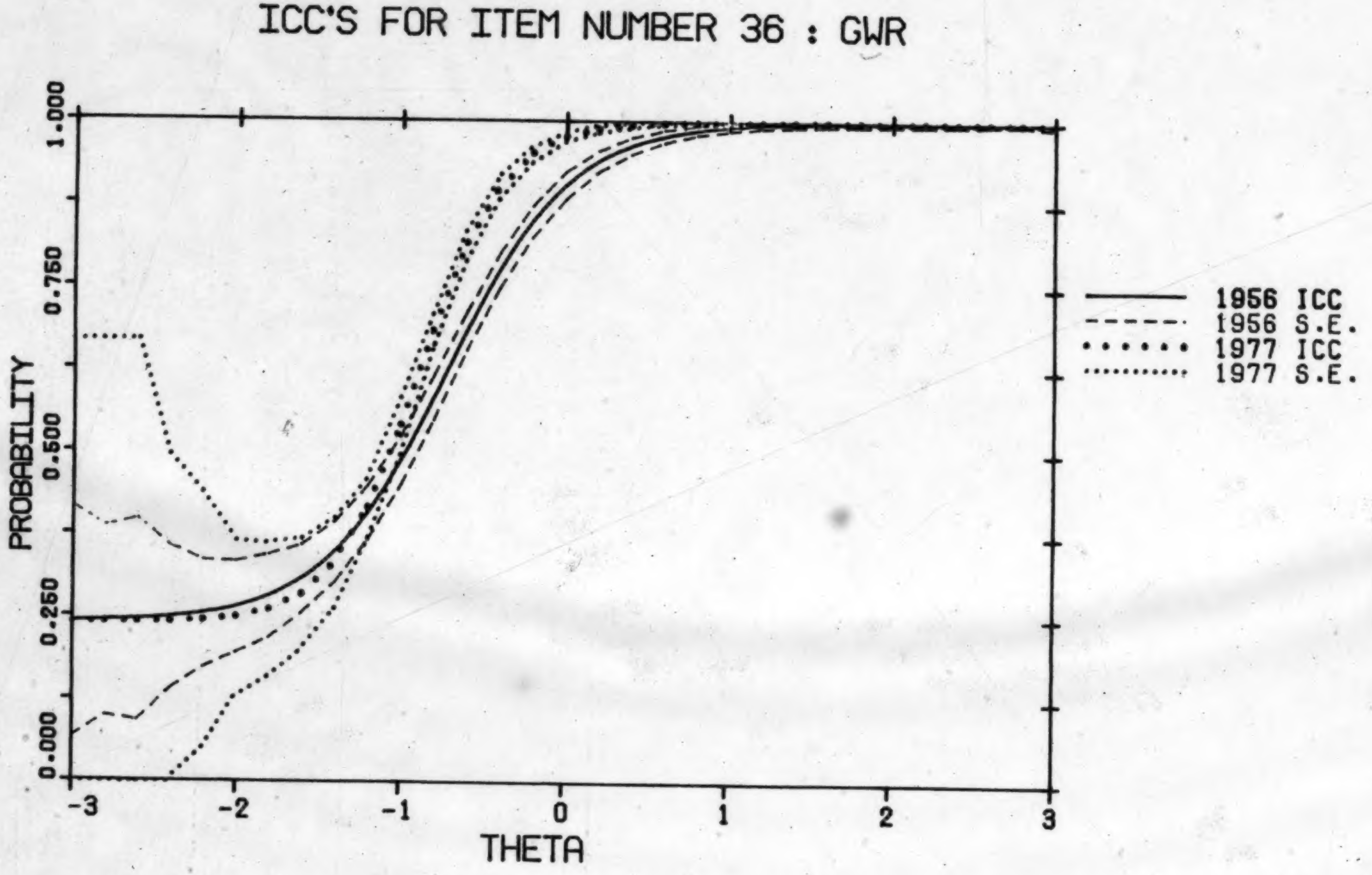

ICC'S FOR ITEM NUMBER 36 : GWR
PROBABILITY
1.000
0.750
0.500
0.250
0.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 78

ICC'S FOR ITEM NUMBER 37 : GWR

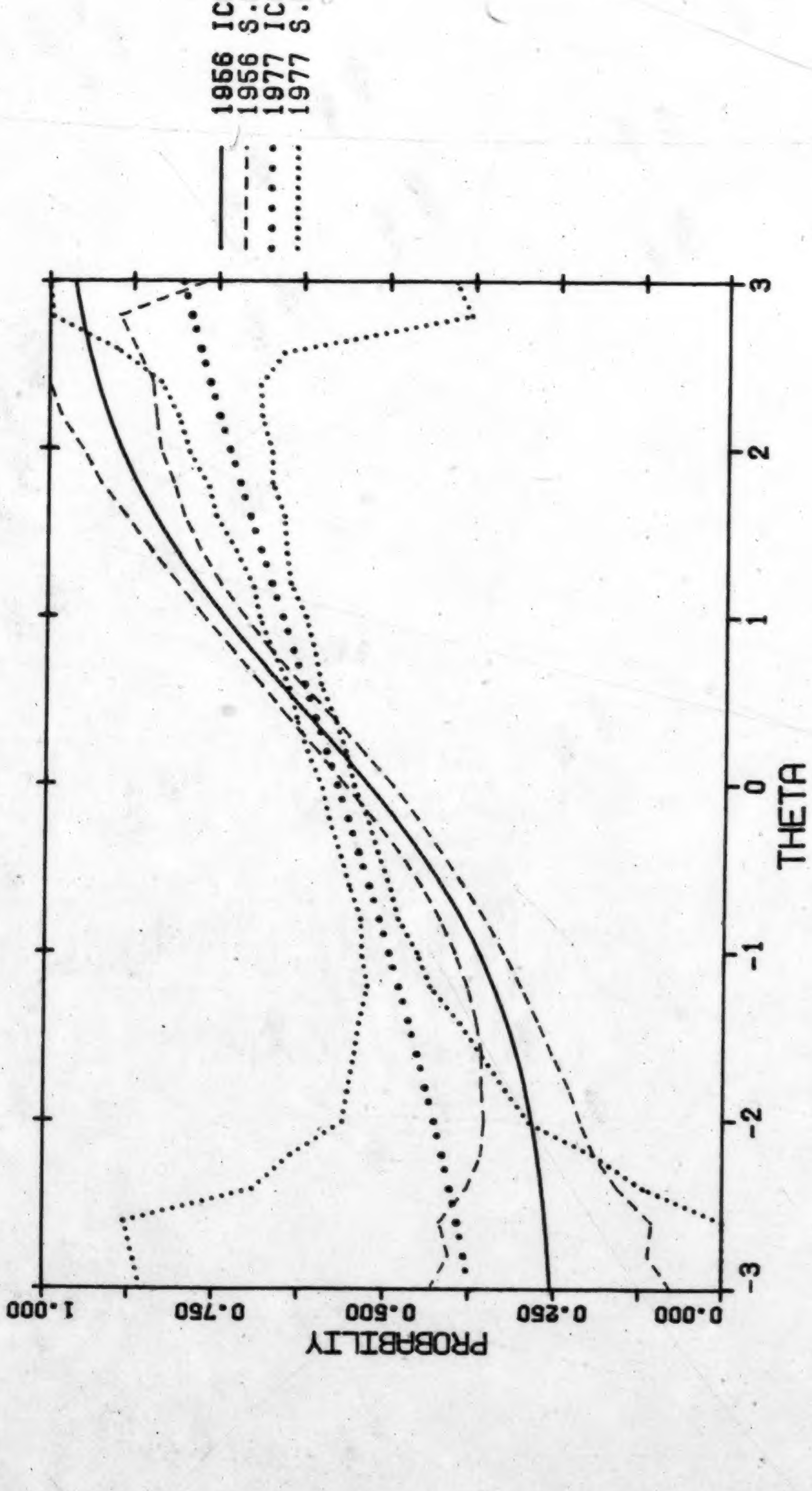


Figure 79

ICC'S FOR ITEM NUMBER 38 : GWR

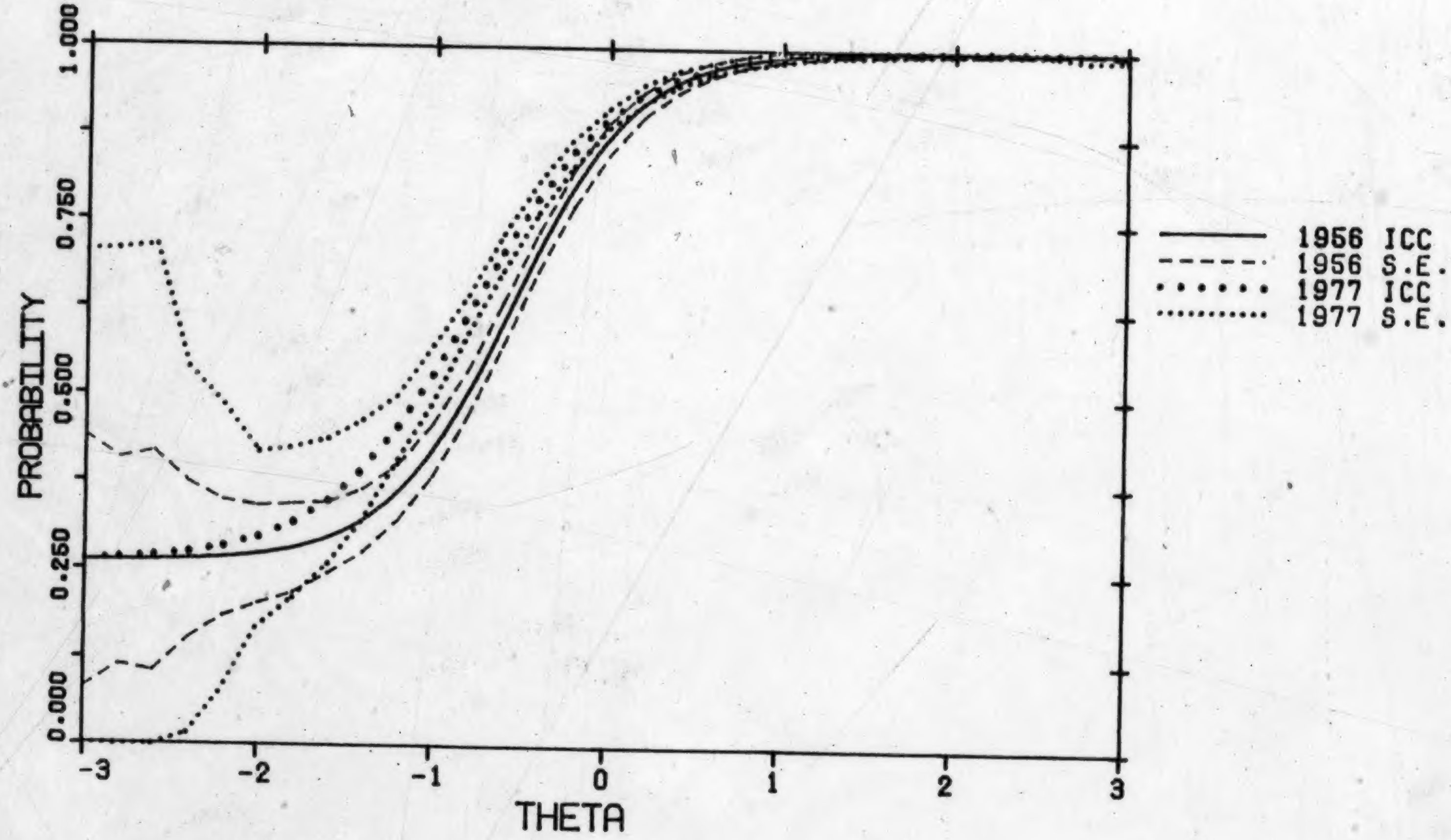


Figure 80

ICC'S FOR ITEM NUMBER 39 : GWR

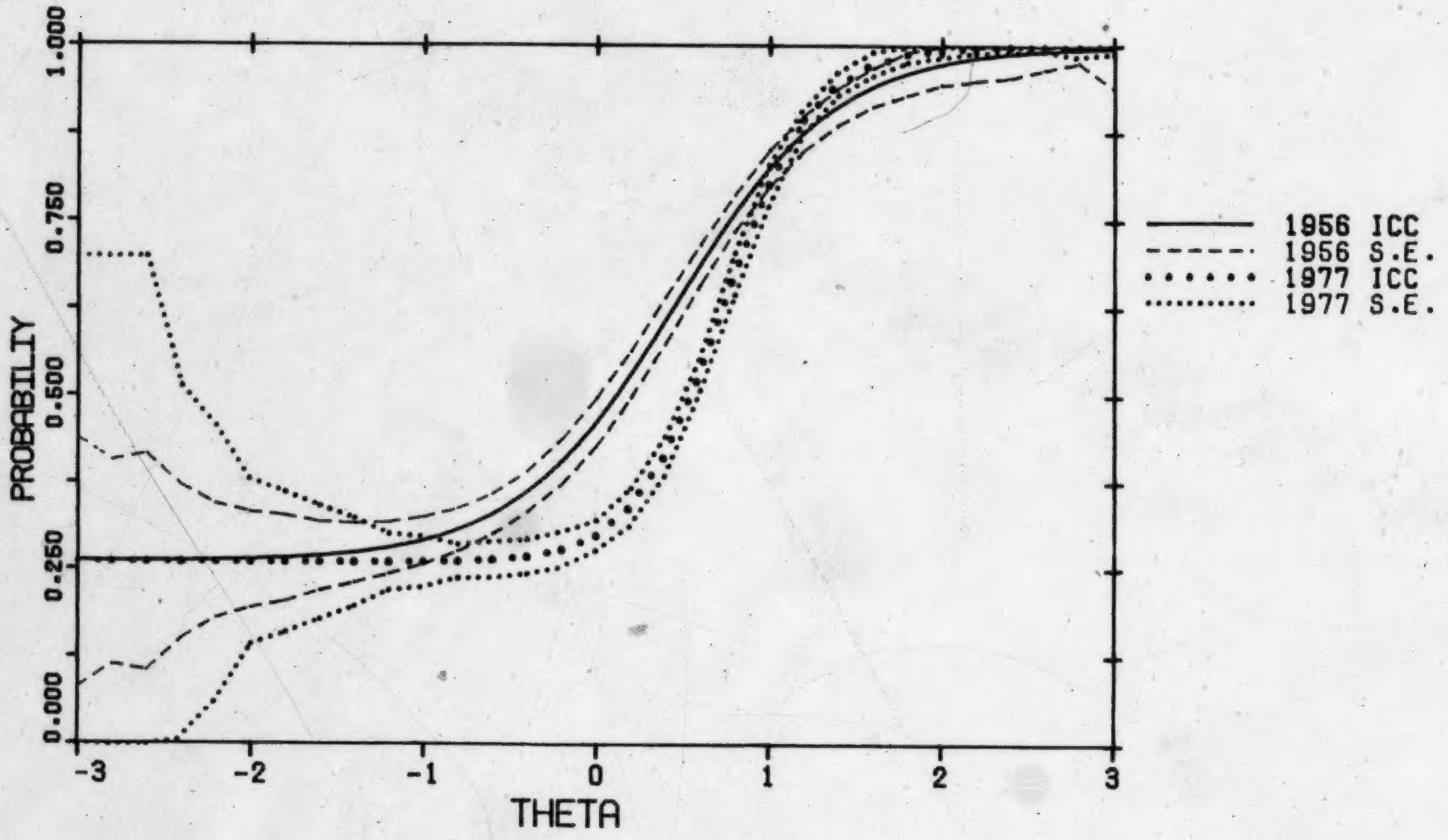


Figure 81

ICC'S FOR ITEM NUMBER 40 : GWR

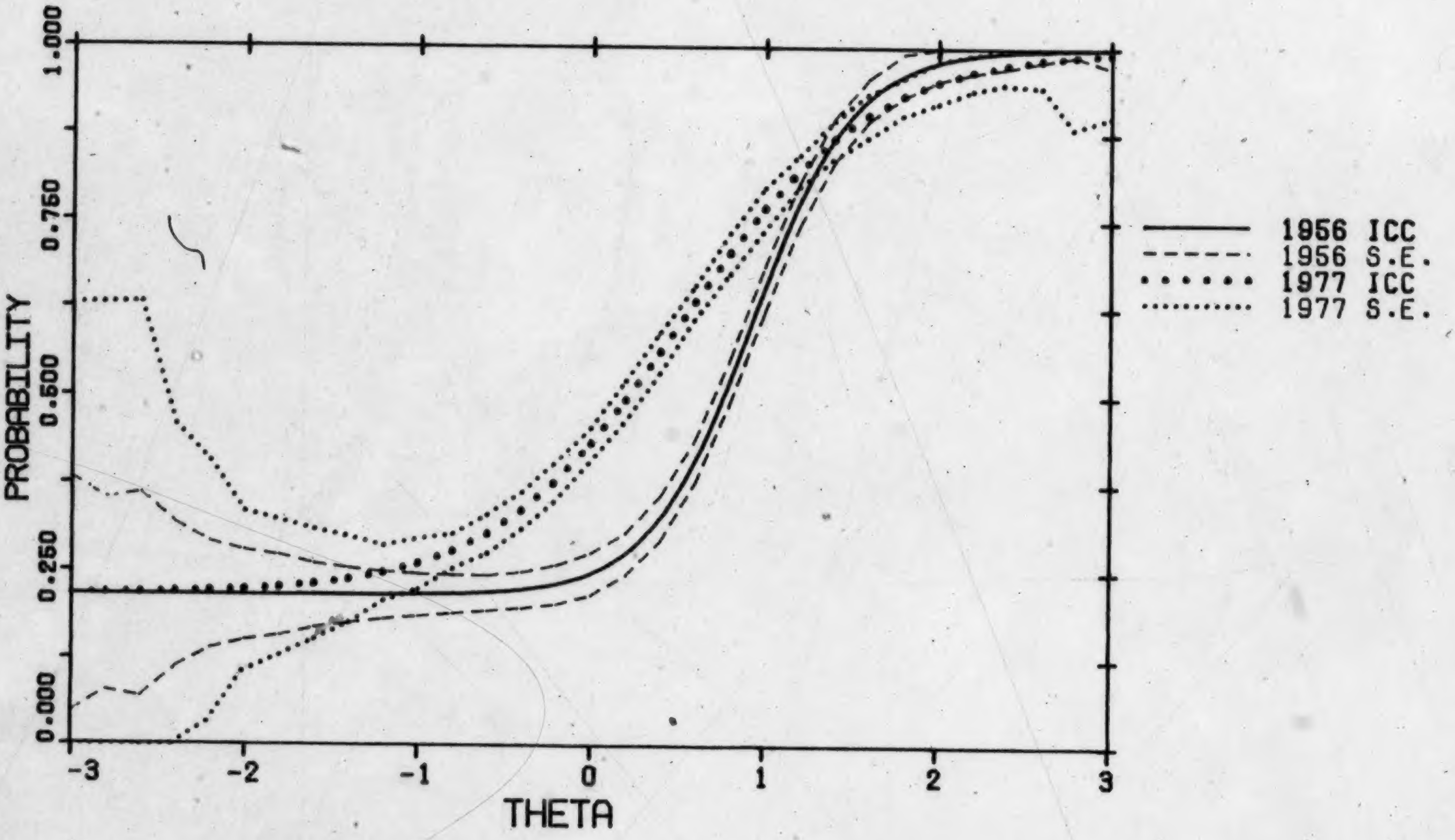


Figure 82

ICC'S FOR ITEM NUMBER 41 : GWR

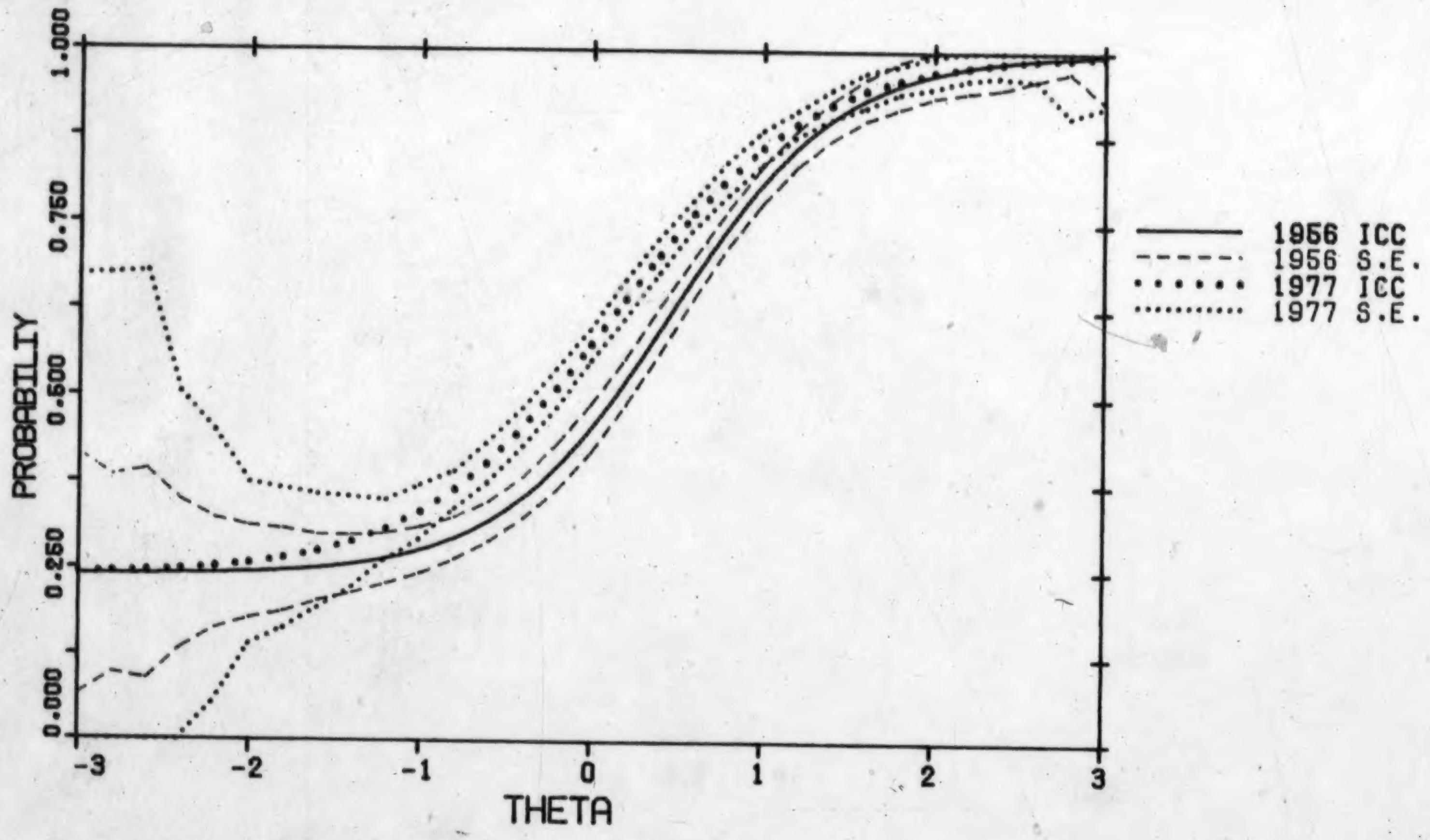


Figure 83

ICC'S FOR ITEM NUMBER 42 : GWR

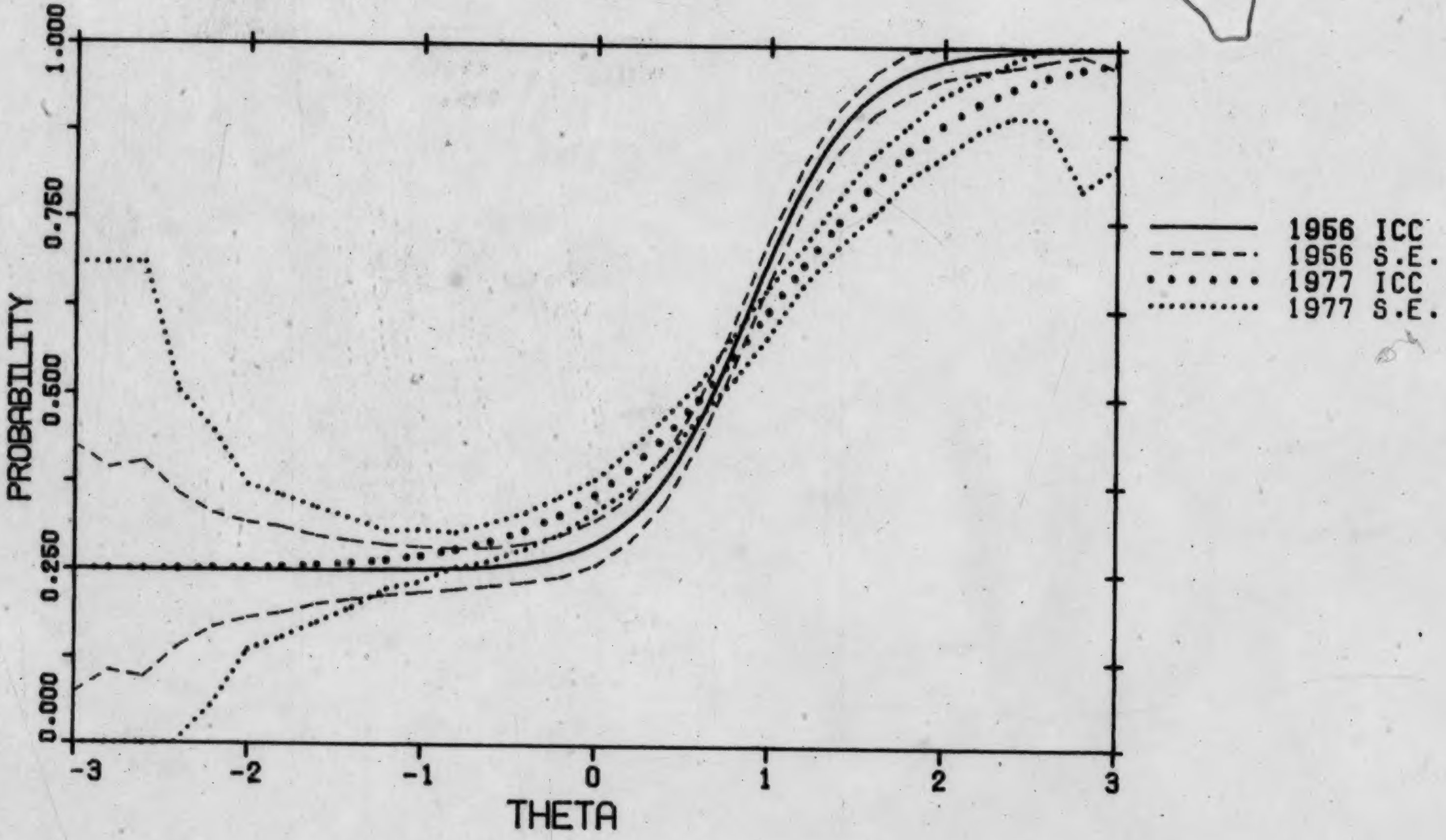


Figure 84

ICC'S FOR ITEM NUMBER 43 : GWR

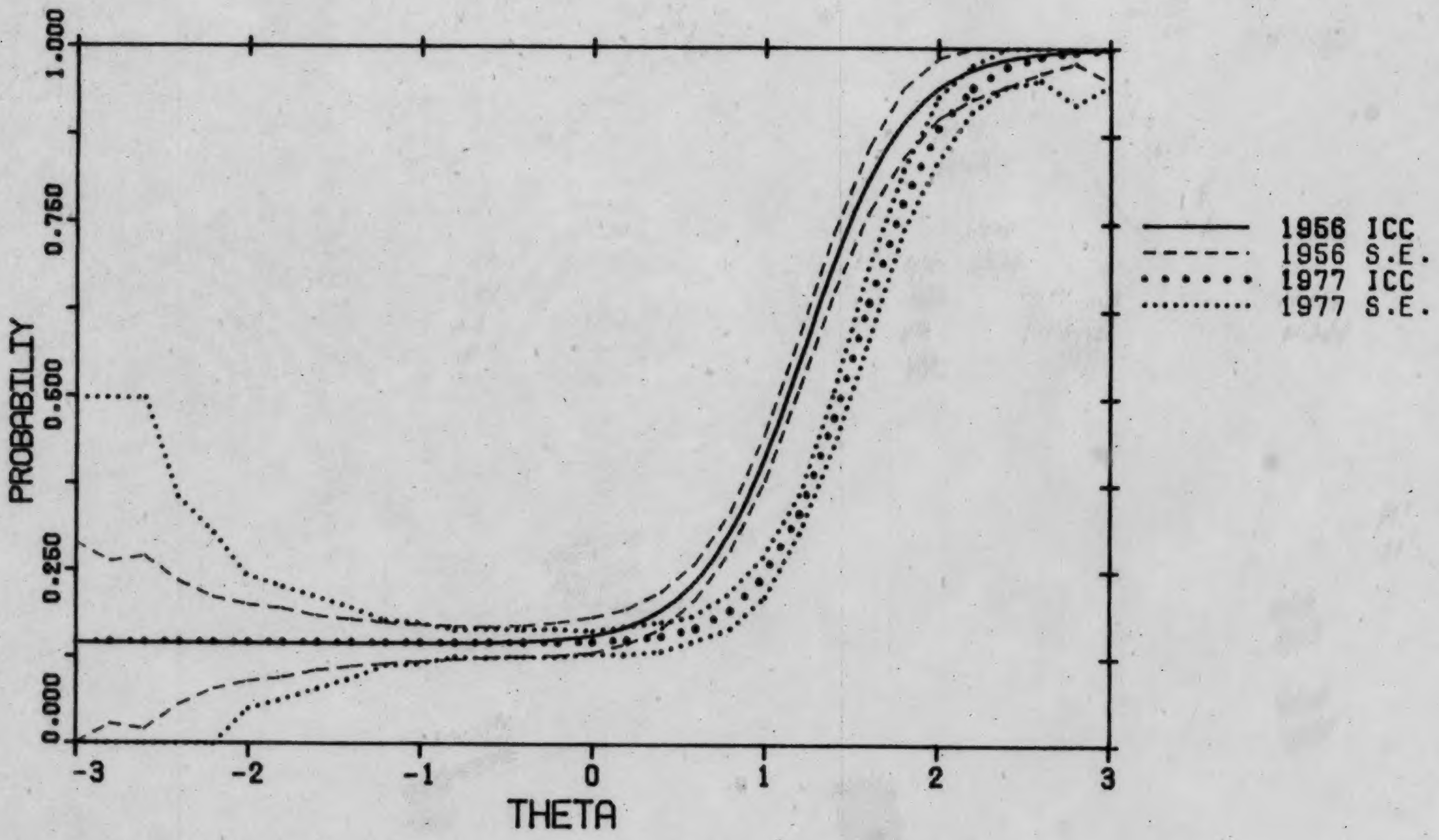


Figure 85

ICC'S FOR ITEM NUMBER 44 : GWR

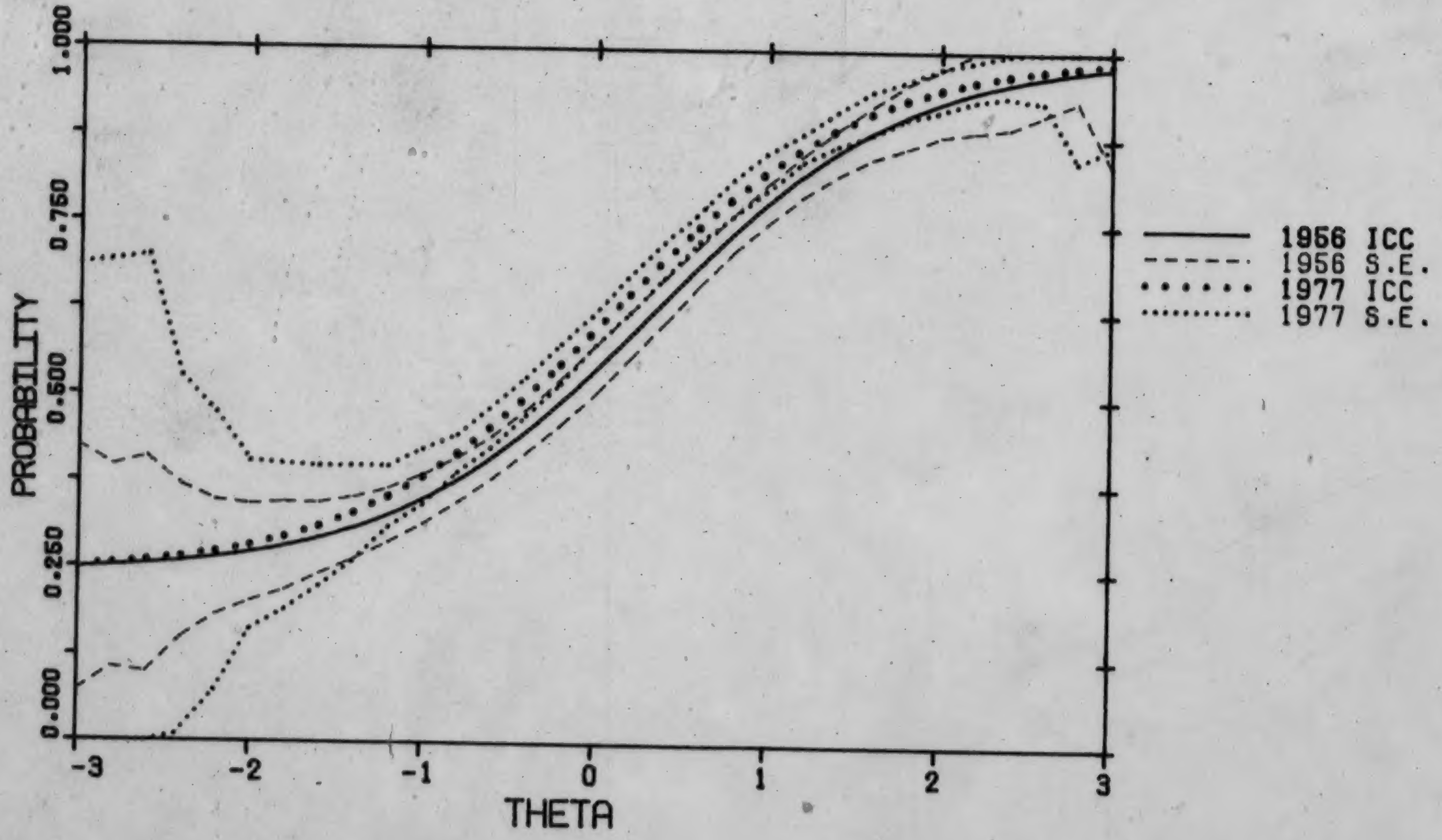


Figure 86

ICC'S FOR ITEM NUMBER 45 : GWR

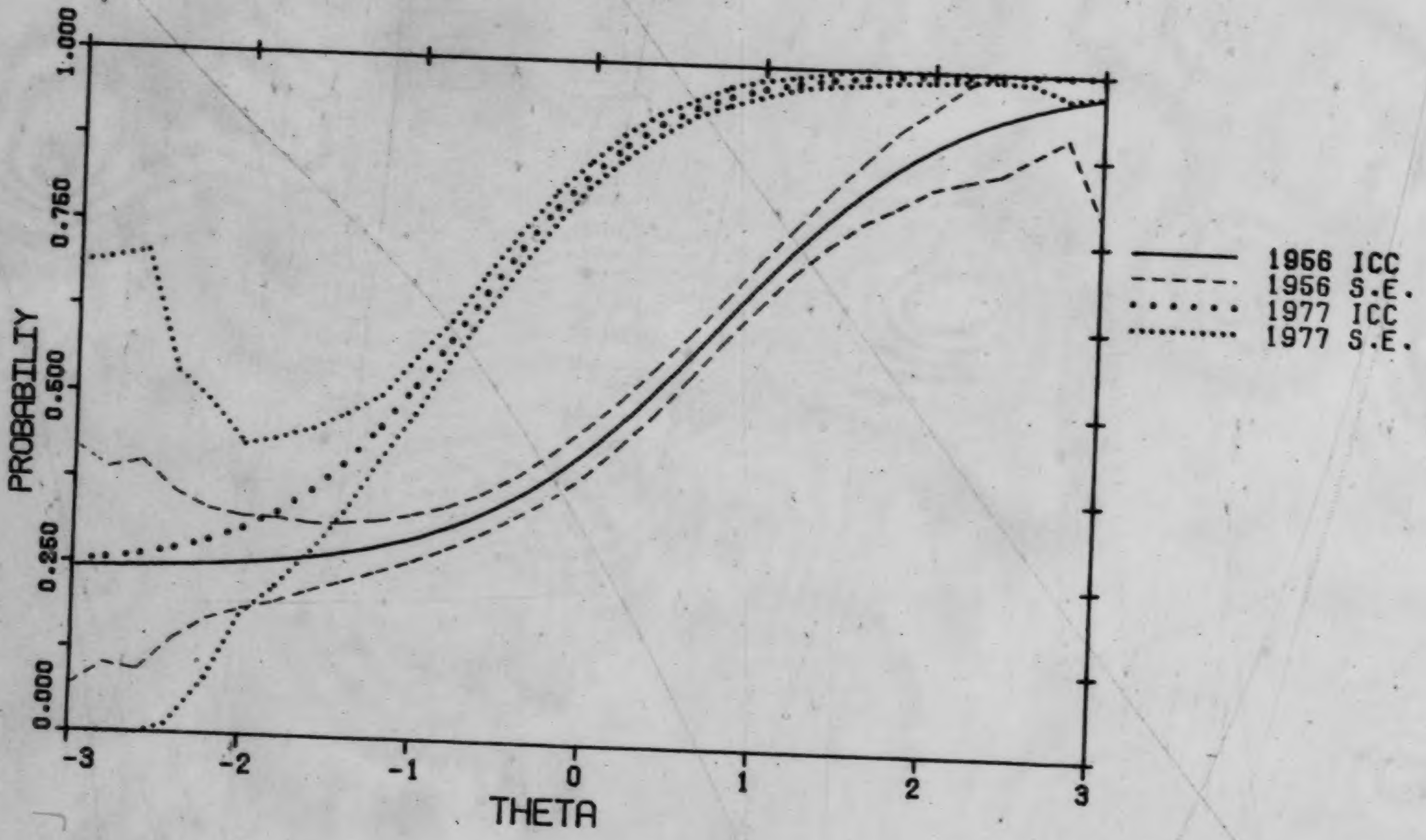


Figure 87

ICC'S FOR ITEM NUMBER 46 : GWR

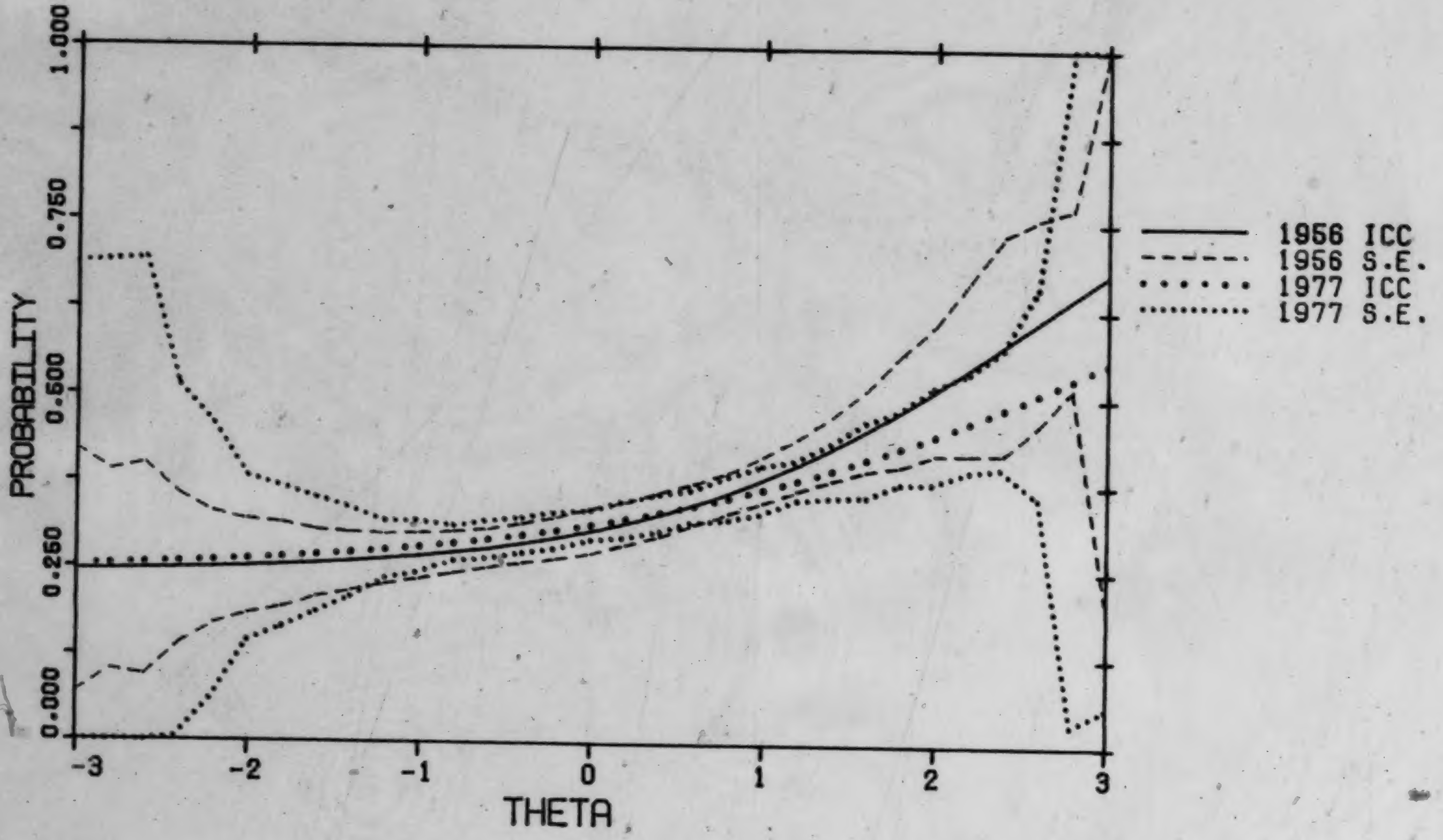
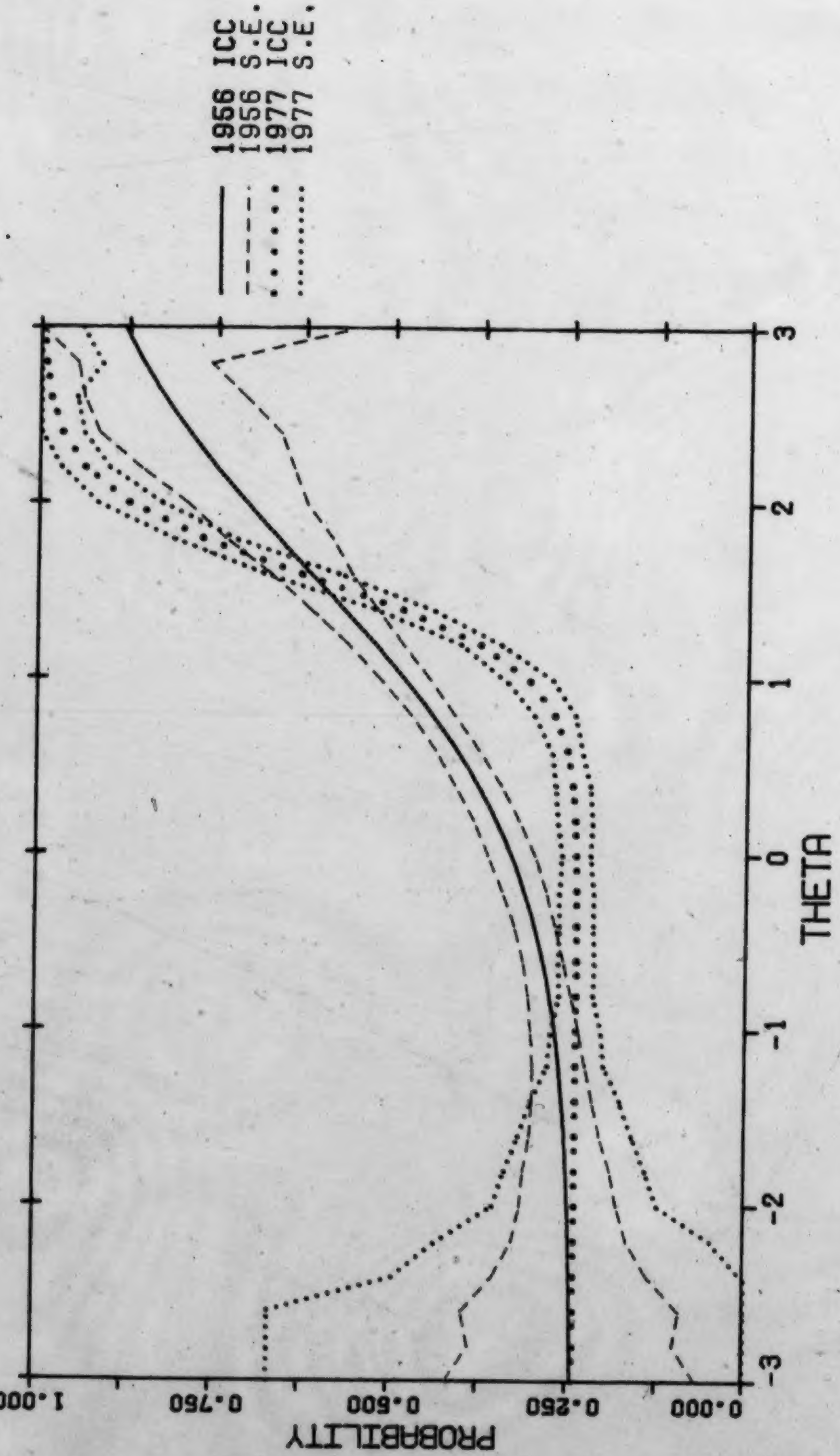


Figure 88

ICC'S FOR ITEM NUMBER 48 : GWR



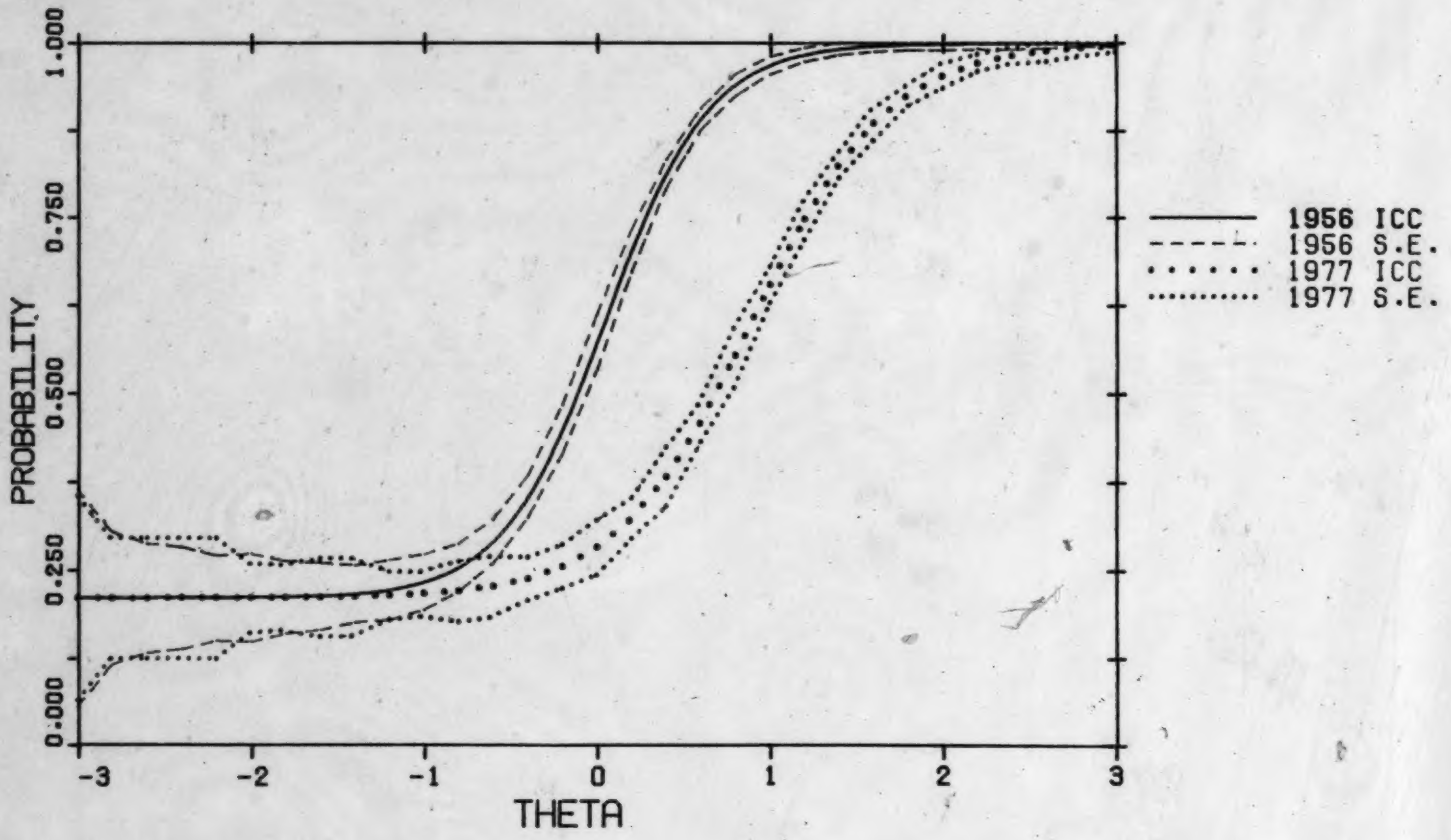


Figure 89

ICC'S FOR ITEM NUMBER 1 : CMM

Figure 90

ICC'S FOR ITEM NUMBER 2 : CMM

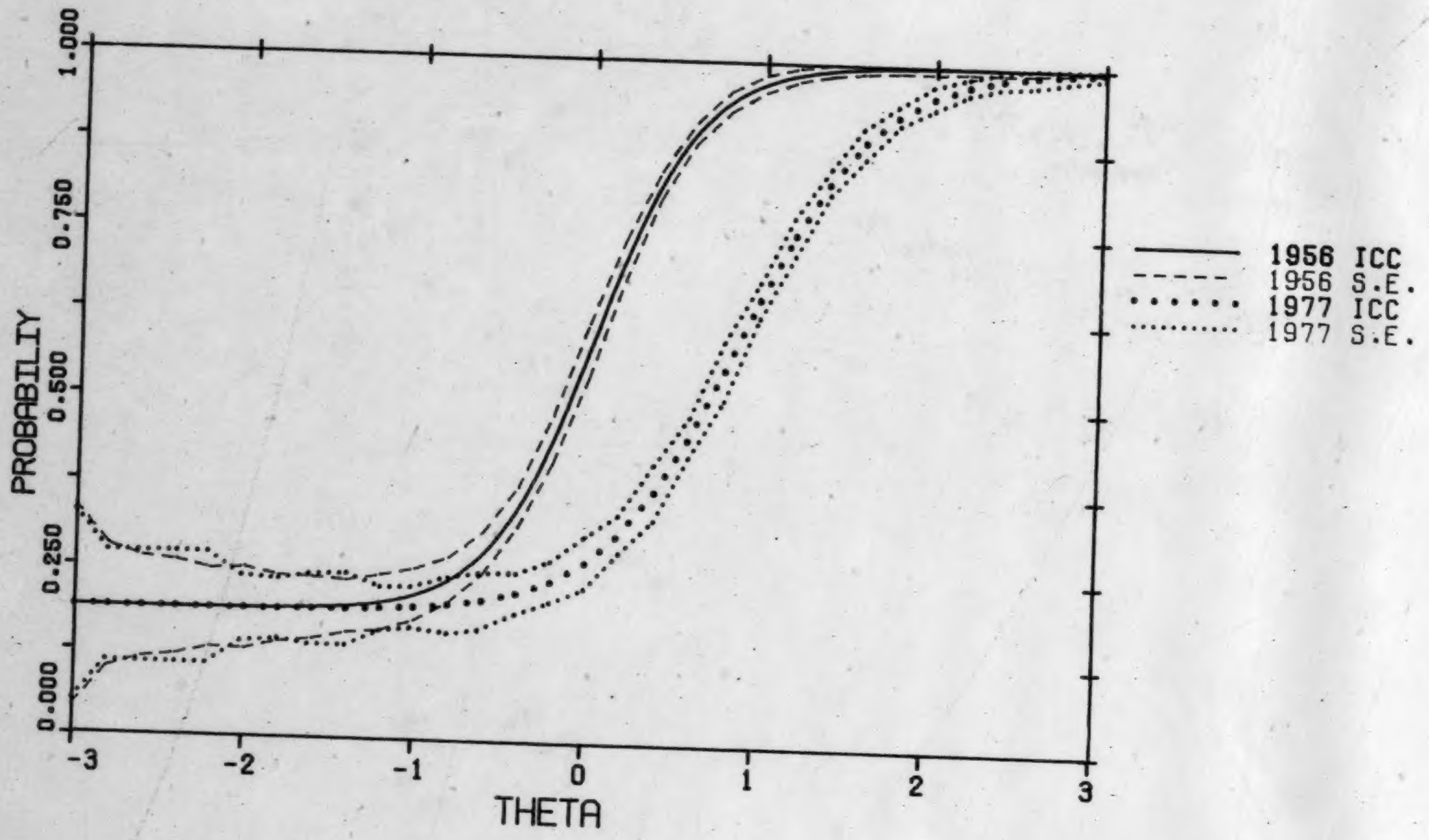


Figure 91

ICC'S FOR ITEM NUMBER 3 : CMM

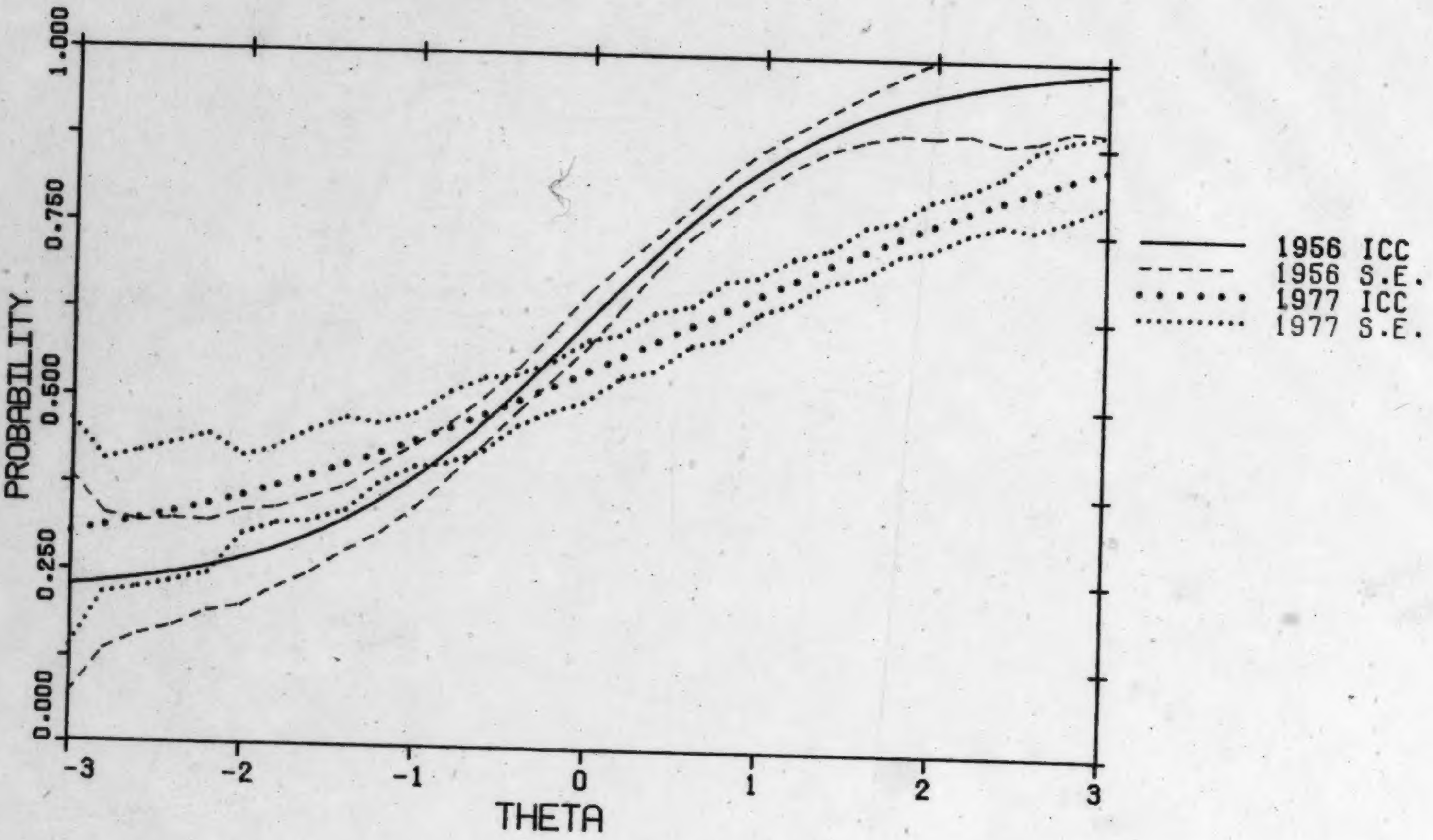


Figure 92

ICC'S FOR ITEM NUMBER 4 : CMM

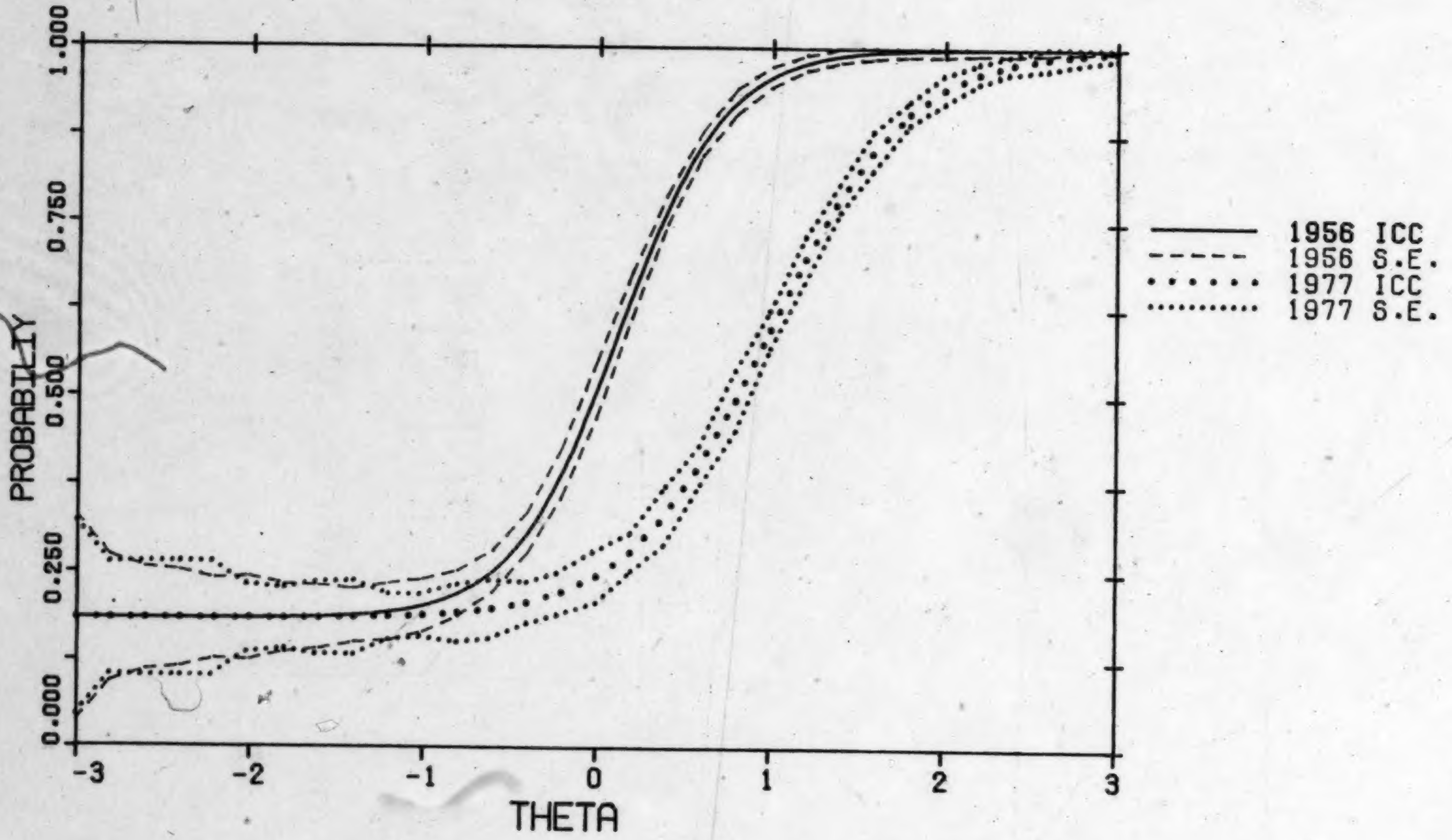


Figure 93

ICC'S FOR ITEM NUMBER 5 : CMM

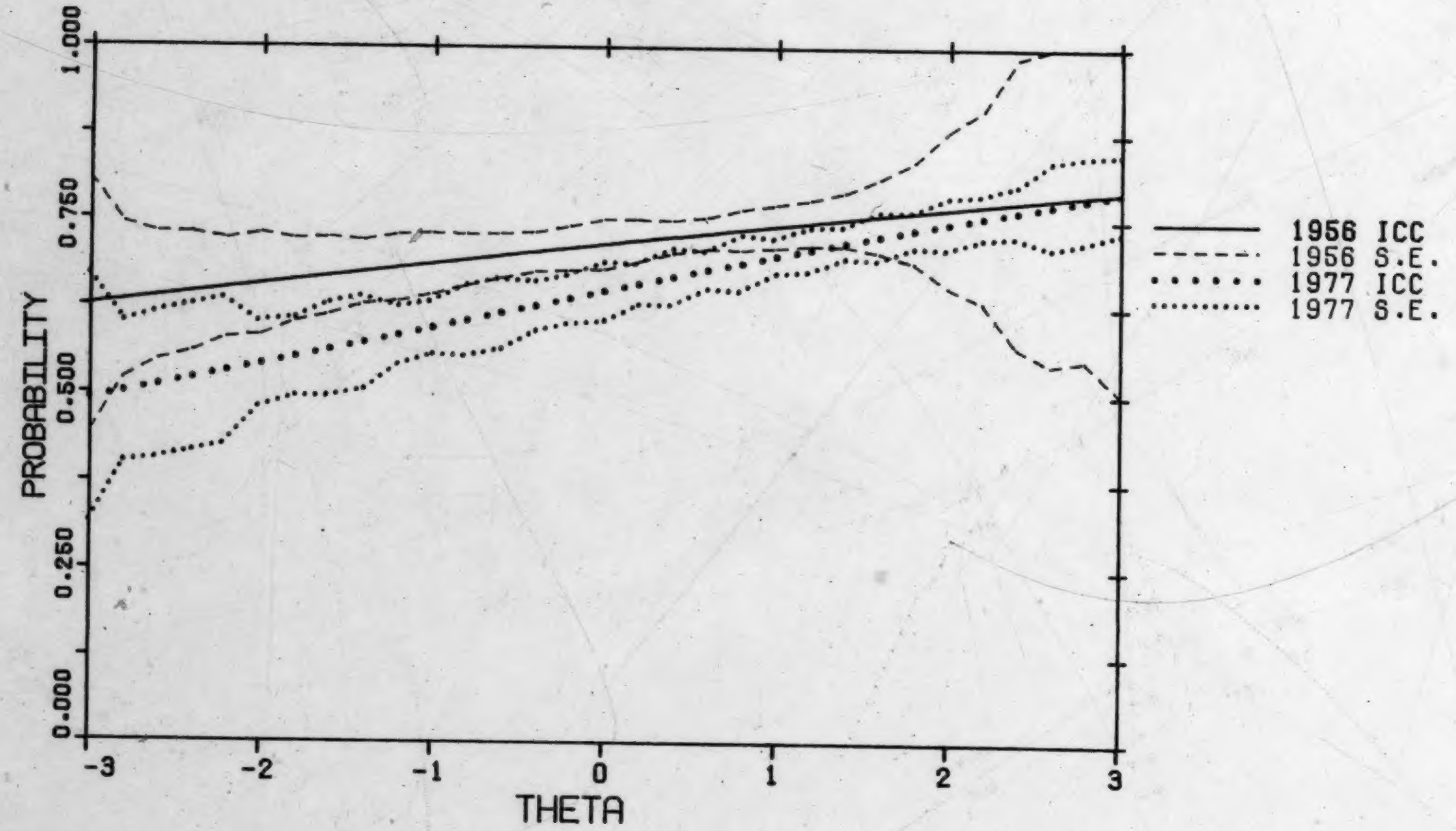


Figure 94

ICC'S FOR ITEM NUMBER 7 : CMM

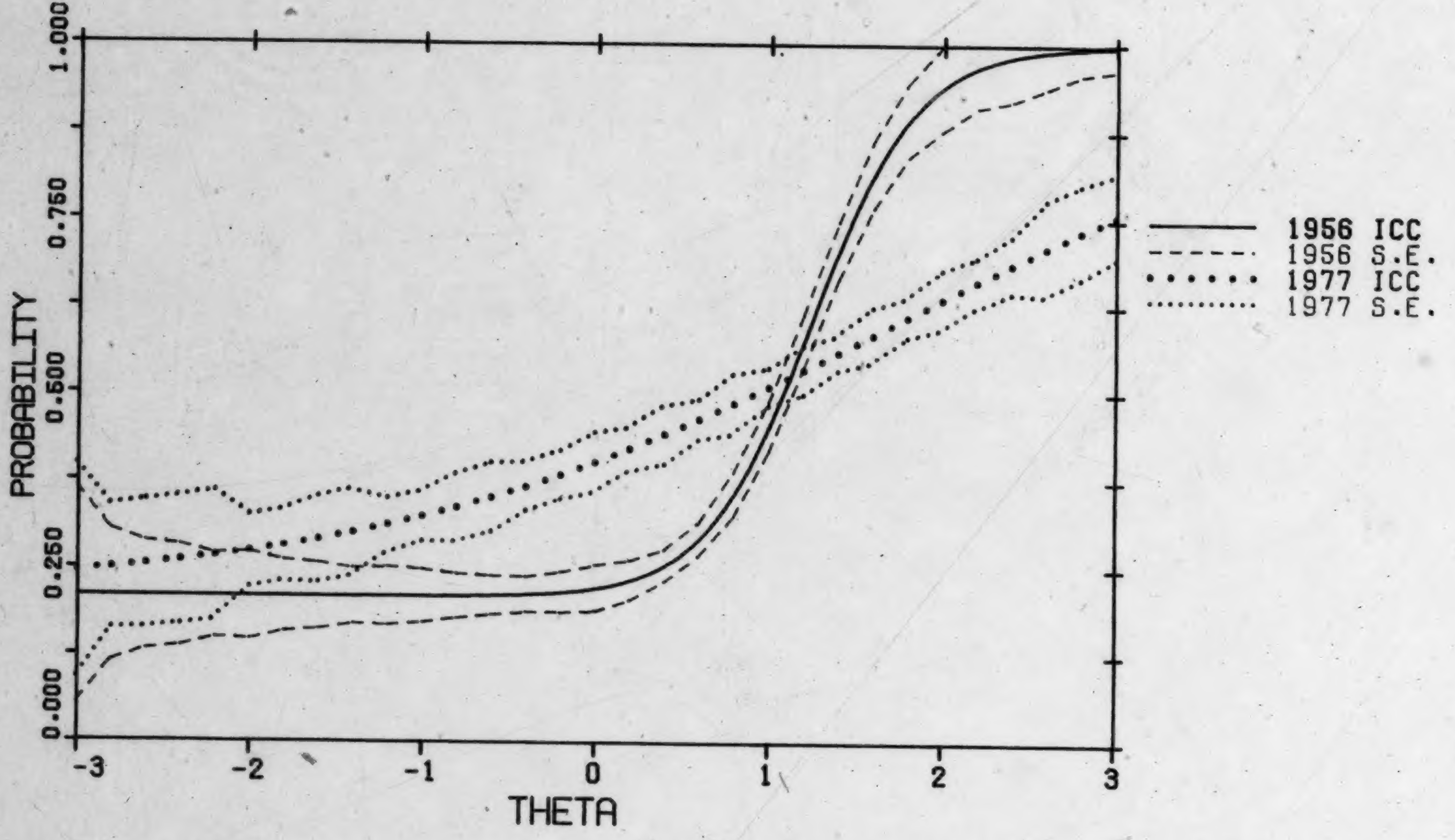


Figure 95

ICC'S FOR ITEM NUMBER 8 : CMM

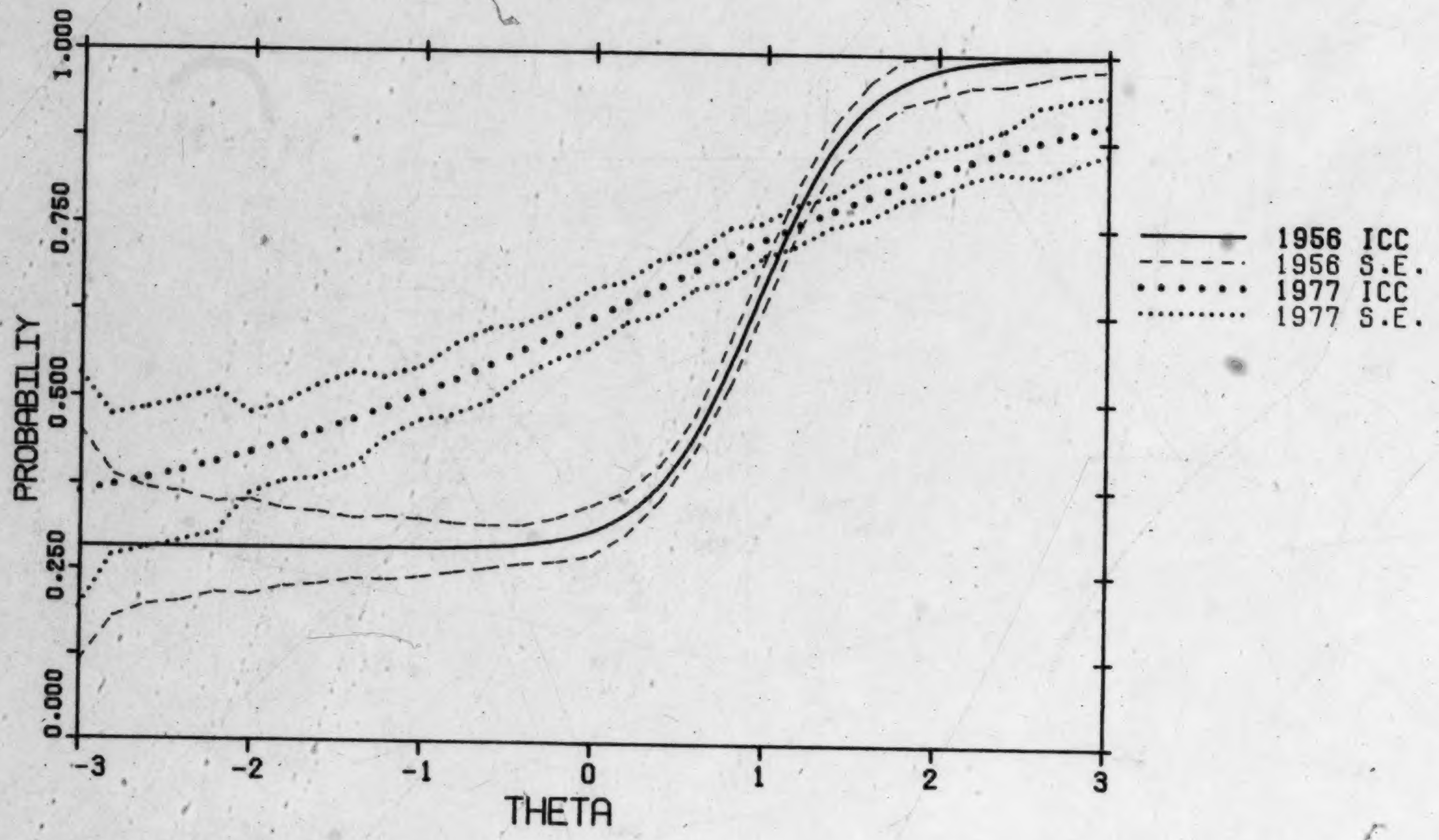


Figure 96

FOR ITEM NUMBER 9 : CMM

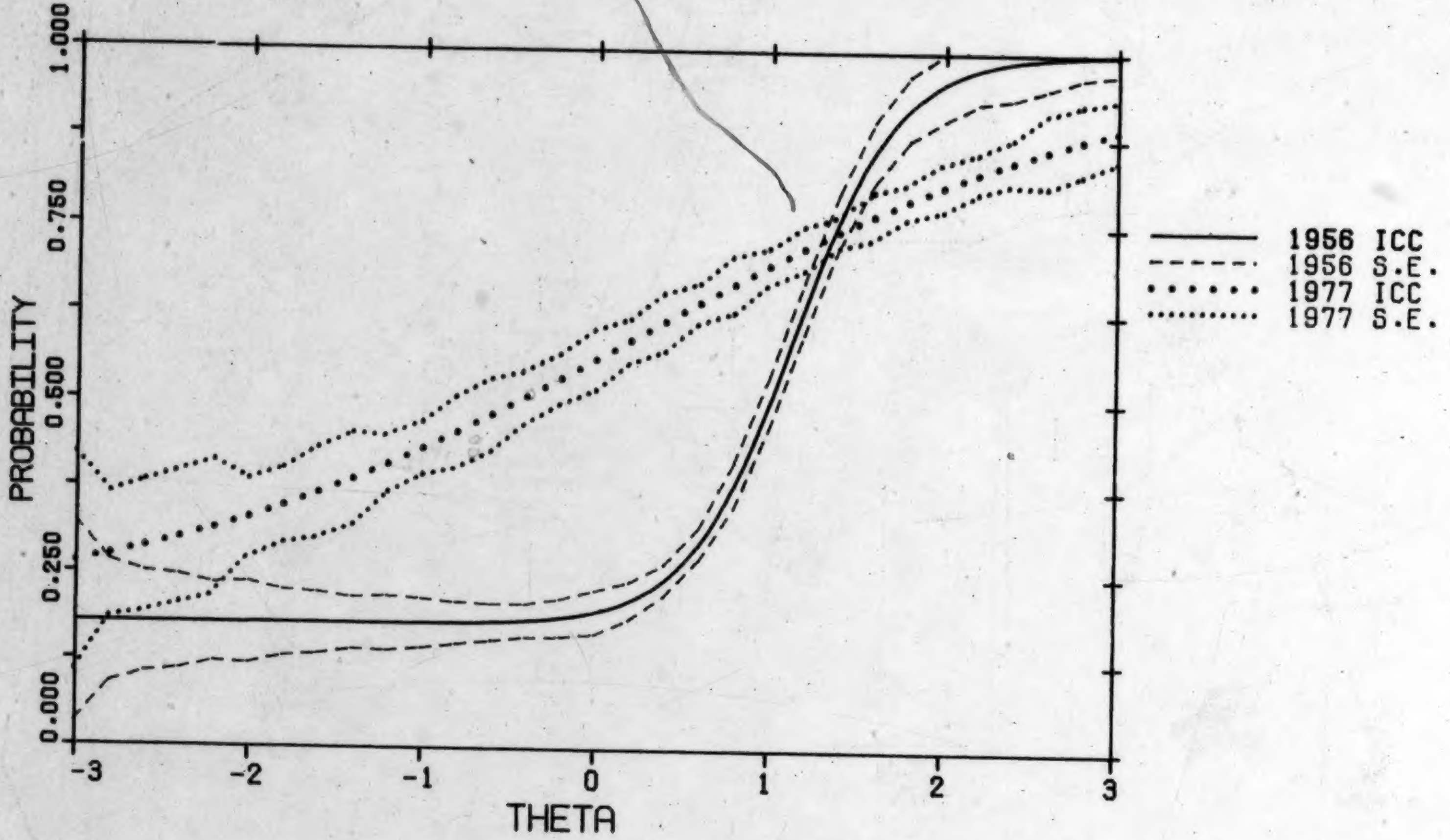


Figure 97

ICC'S FOR ITEM NUMBER 10 : CMM

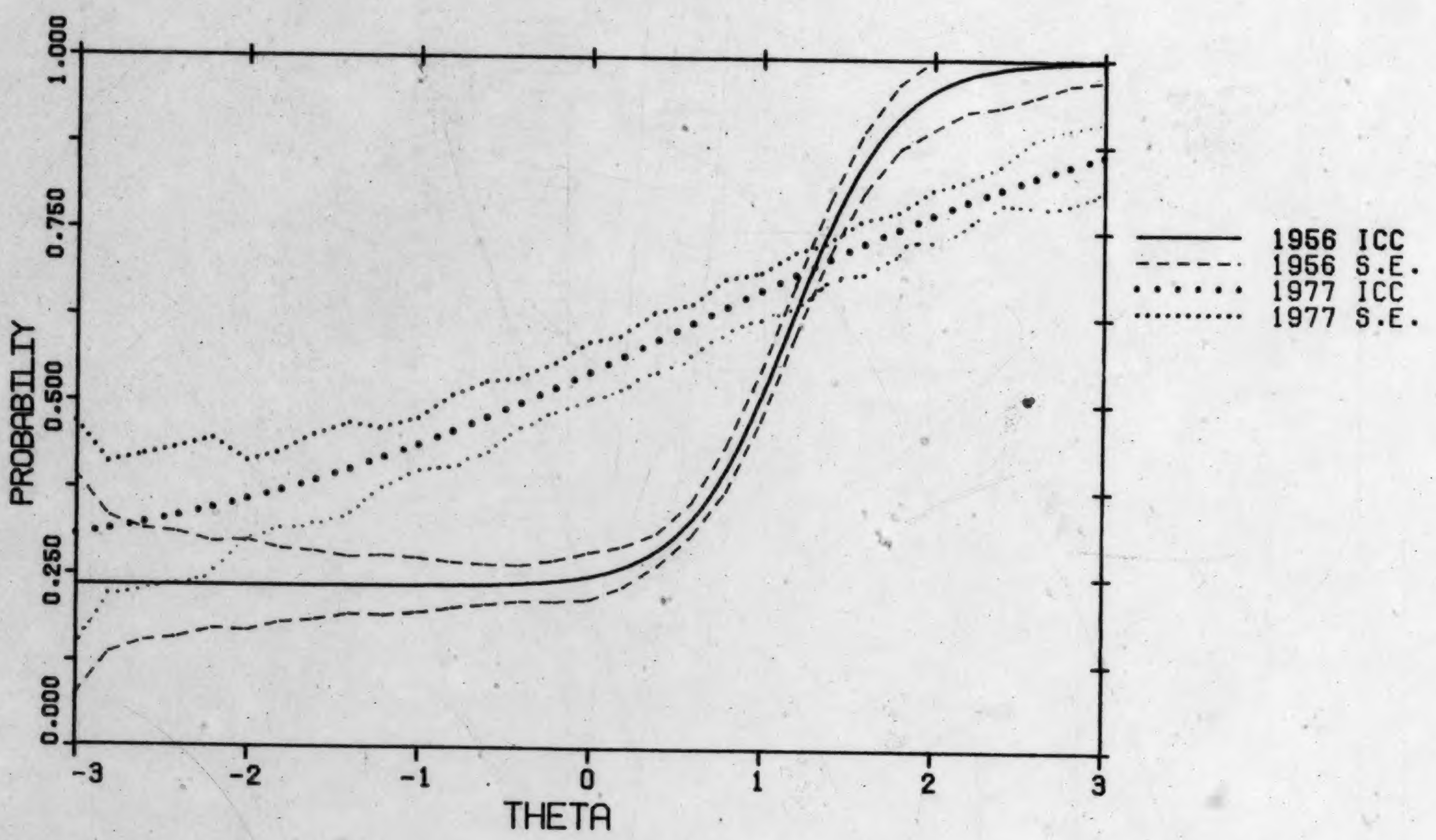


Figure 98

ICC'S FOR ITEM NUMBER 11 : CMM

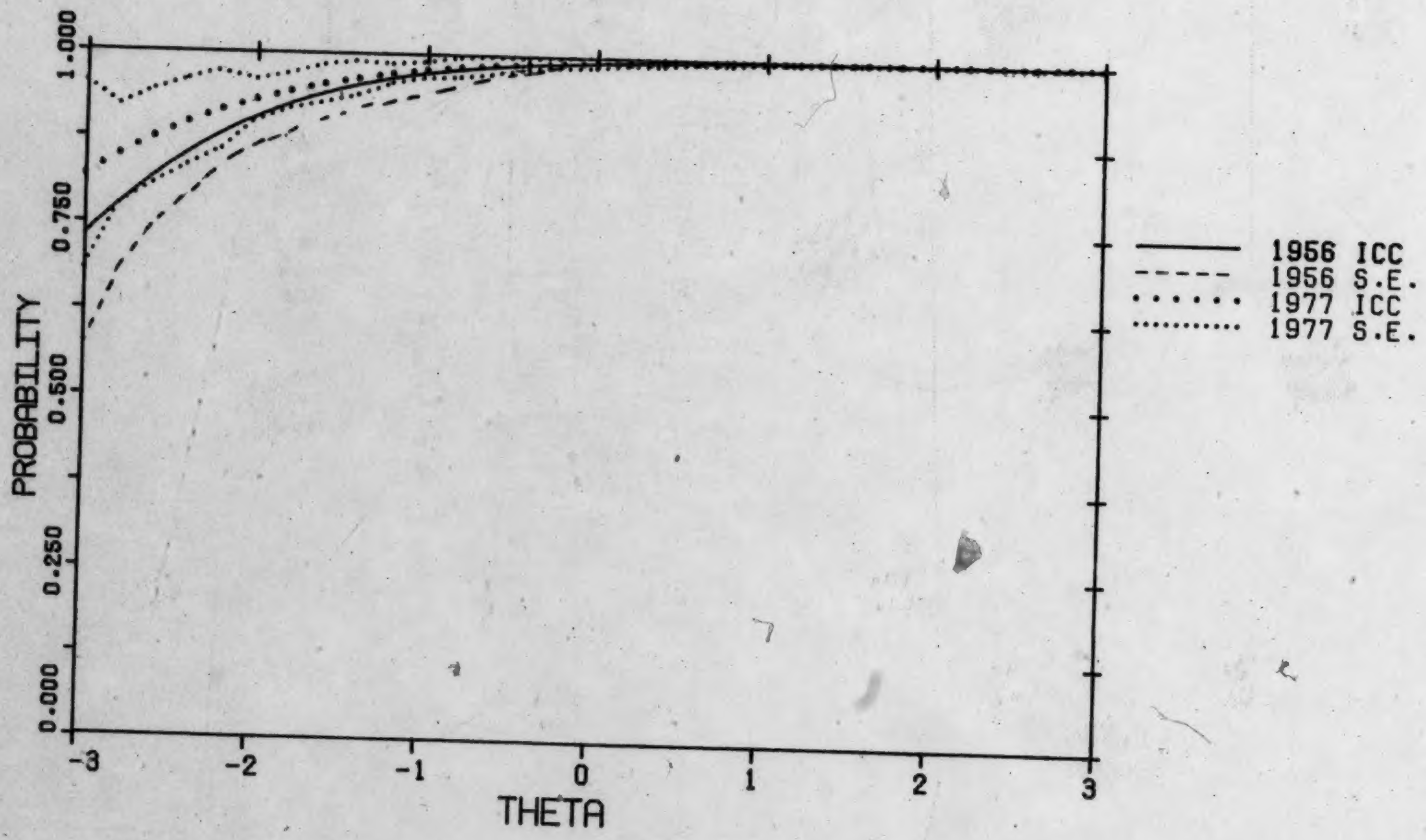


Figure 99

ICC'S FOR ITEM NUMBER 12 : CMM

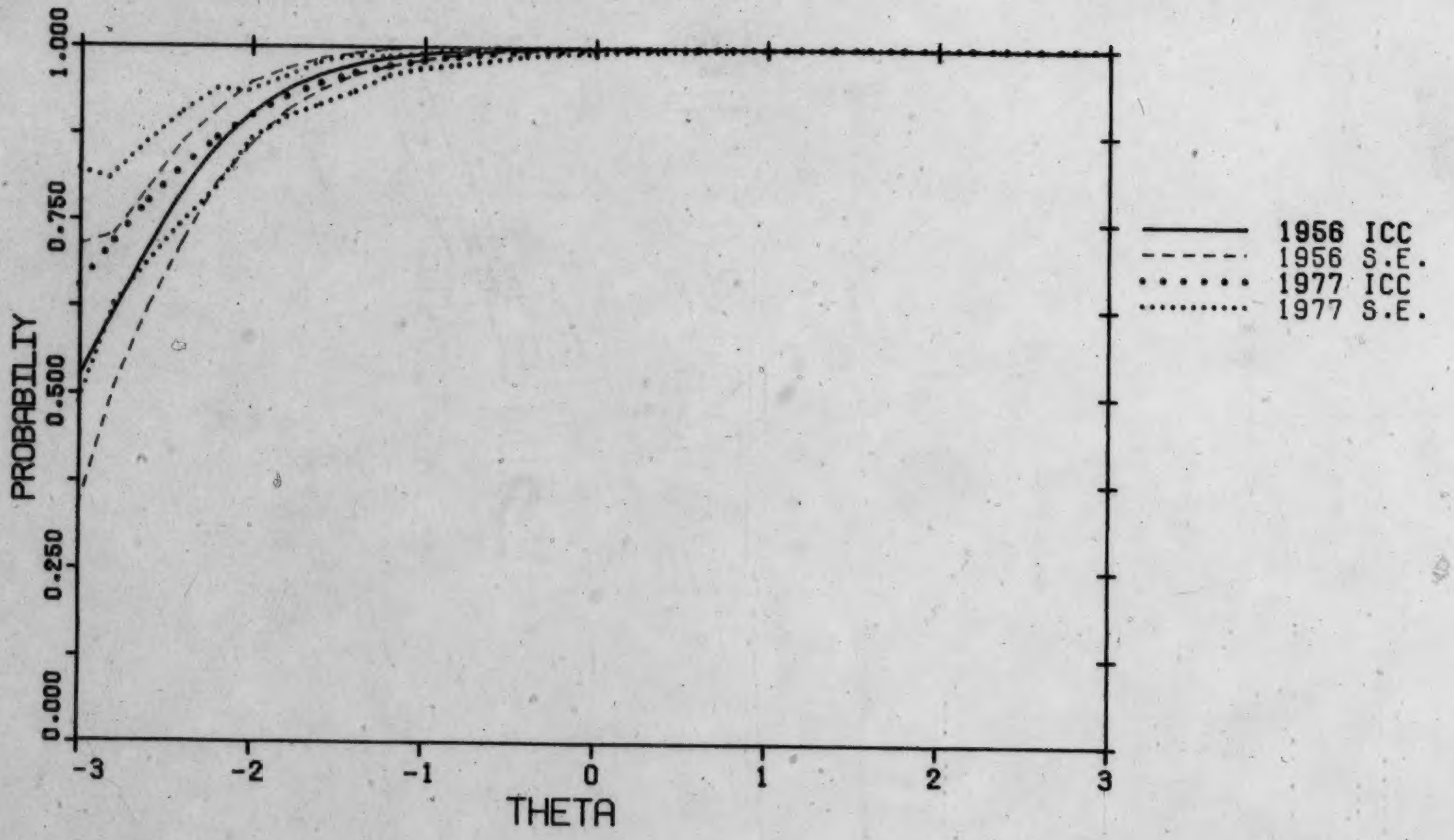


Figure 100

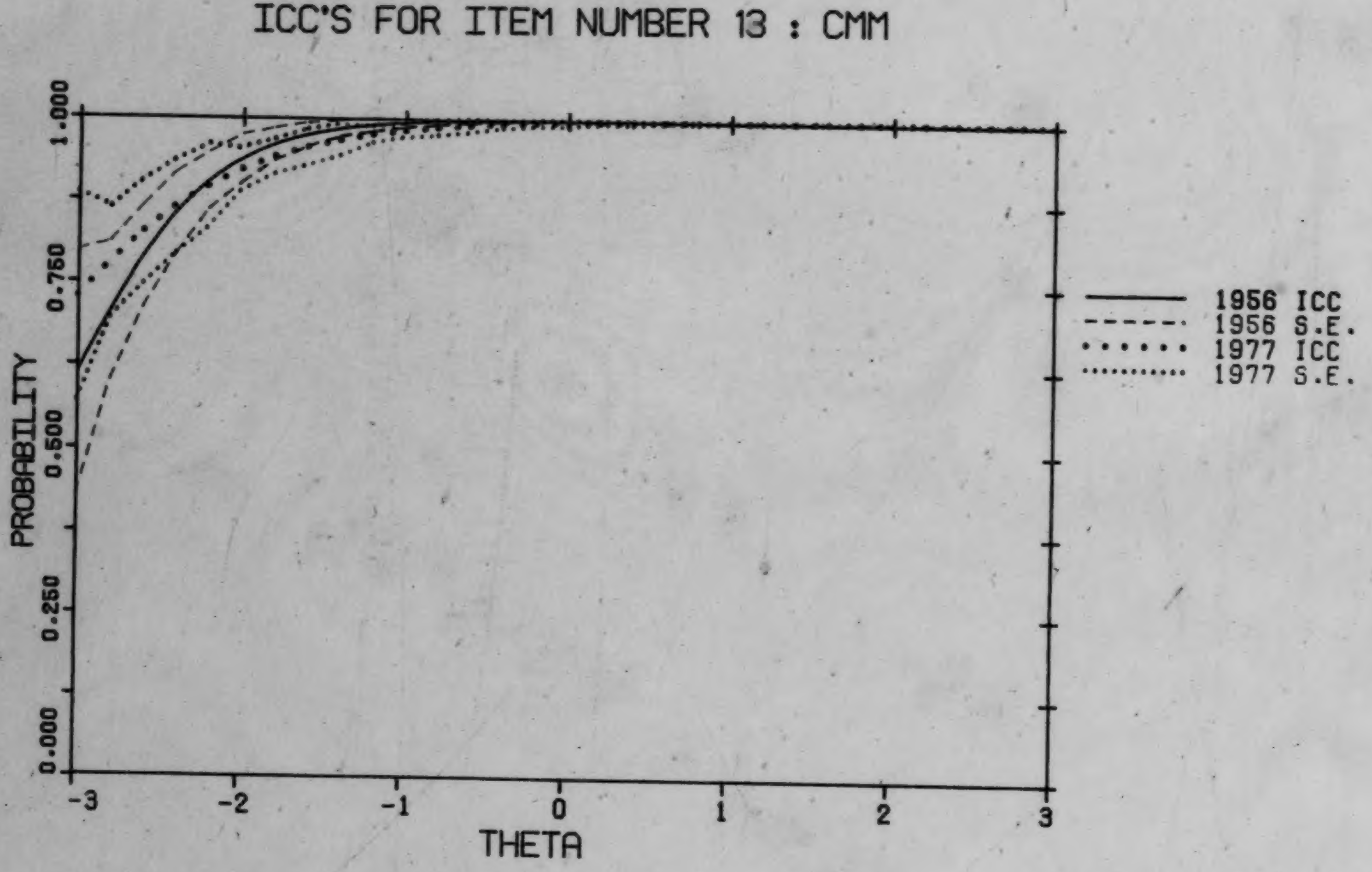

ICC'S FOR ITEM NUMBER 13 : CMM
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 101

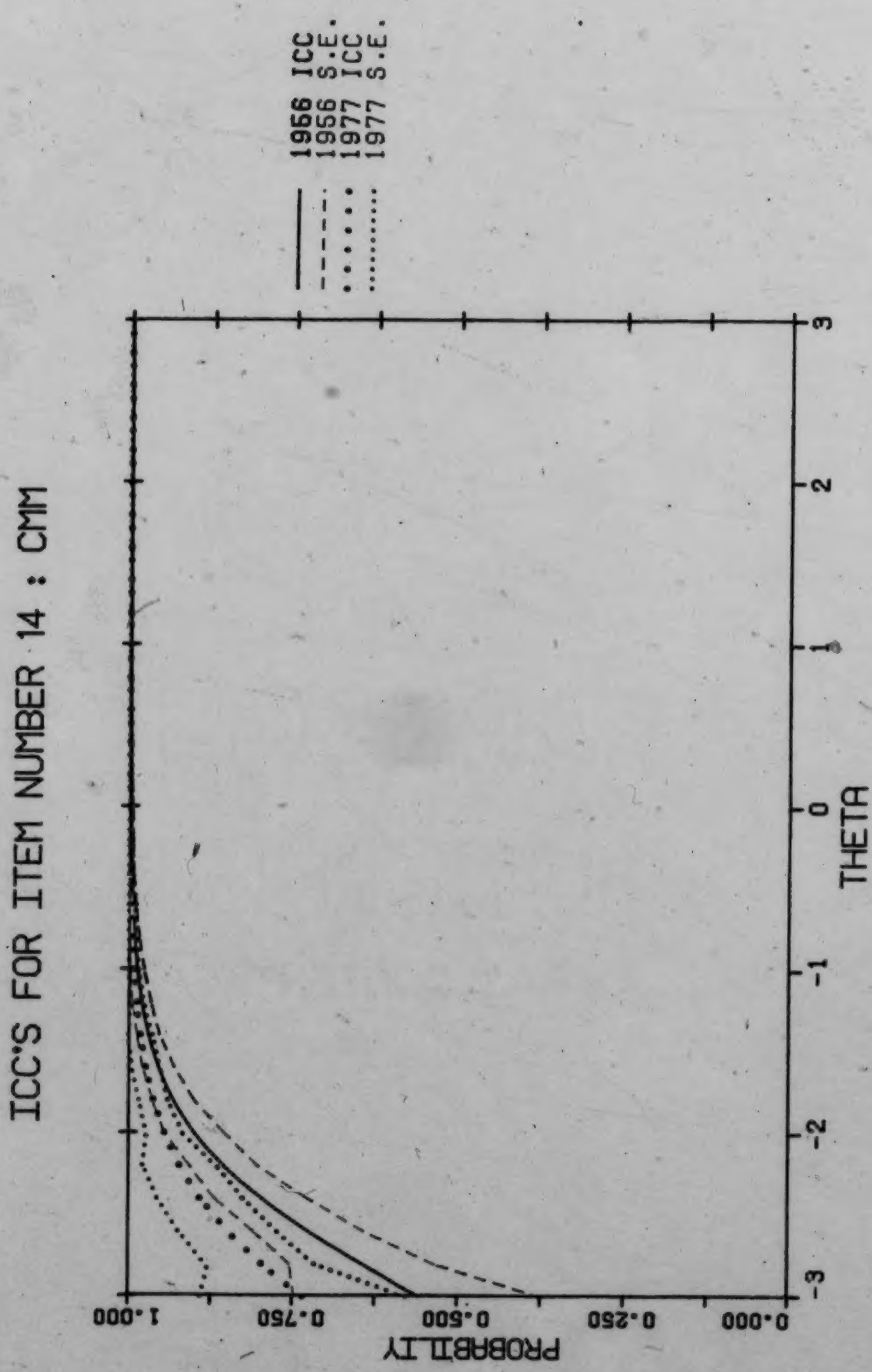
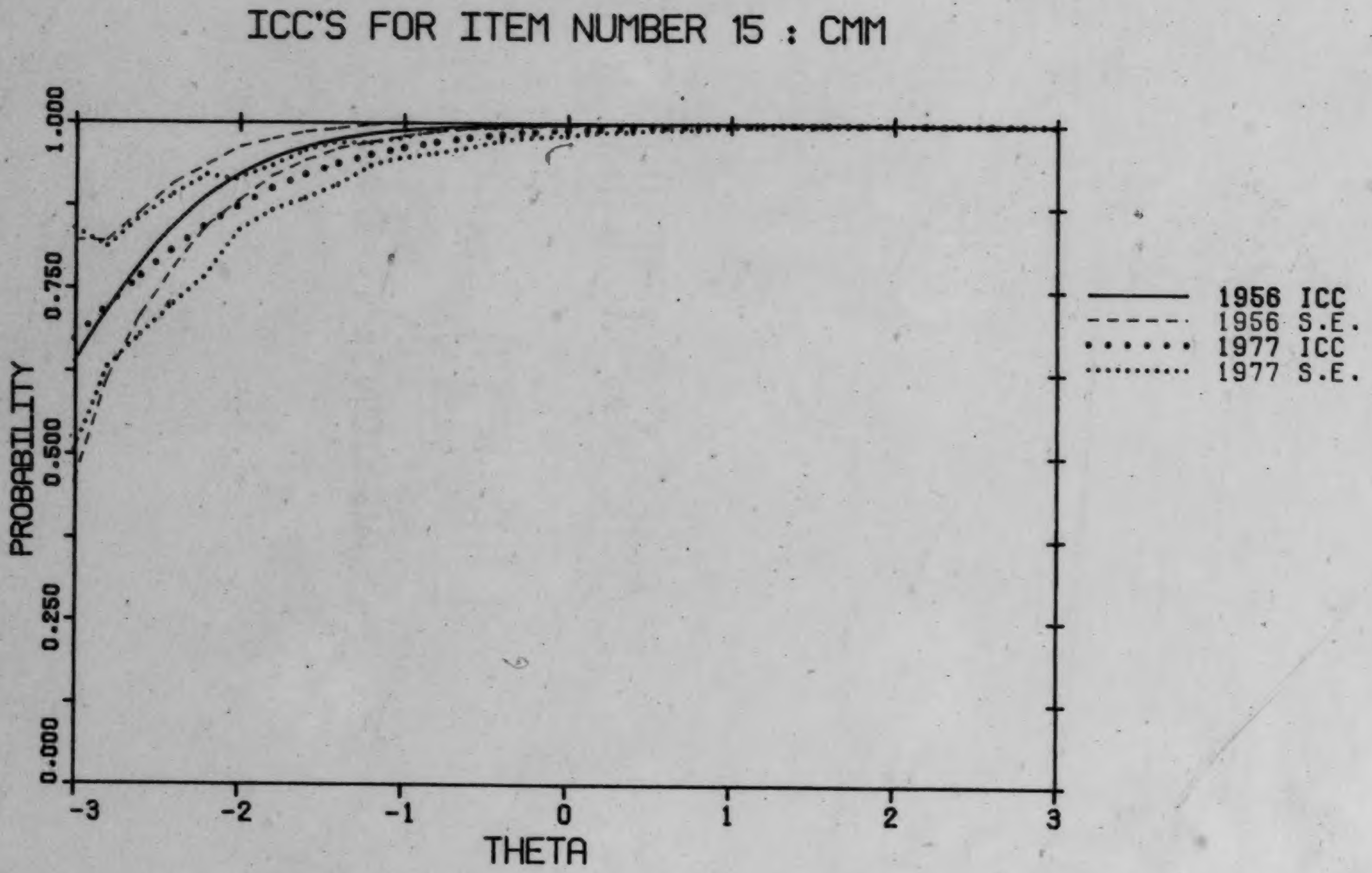

ICC'S FOR ITEM NUMBER 14 : CMM
PROBABILITY
THETA
0.000
0.250
0.500
0.750
1.000
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 102



ICC'S FOR ITEM NUMBER 15 : CMM

Figure 103

ICC'S FOR ITEM NUMBER 16 : CMM

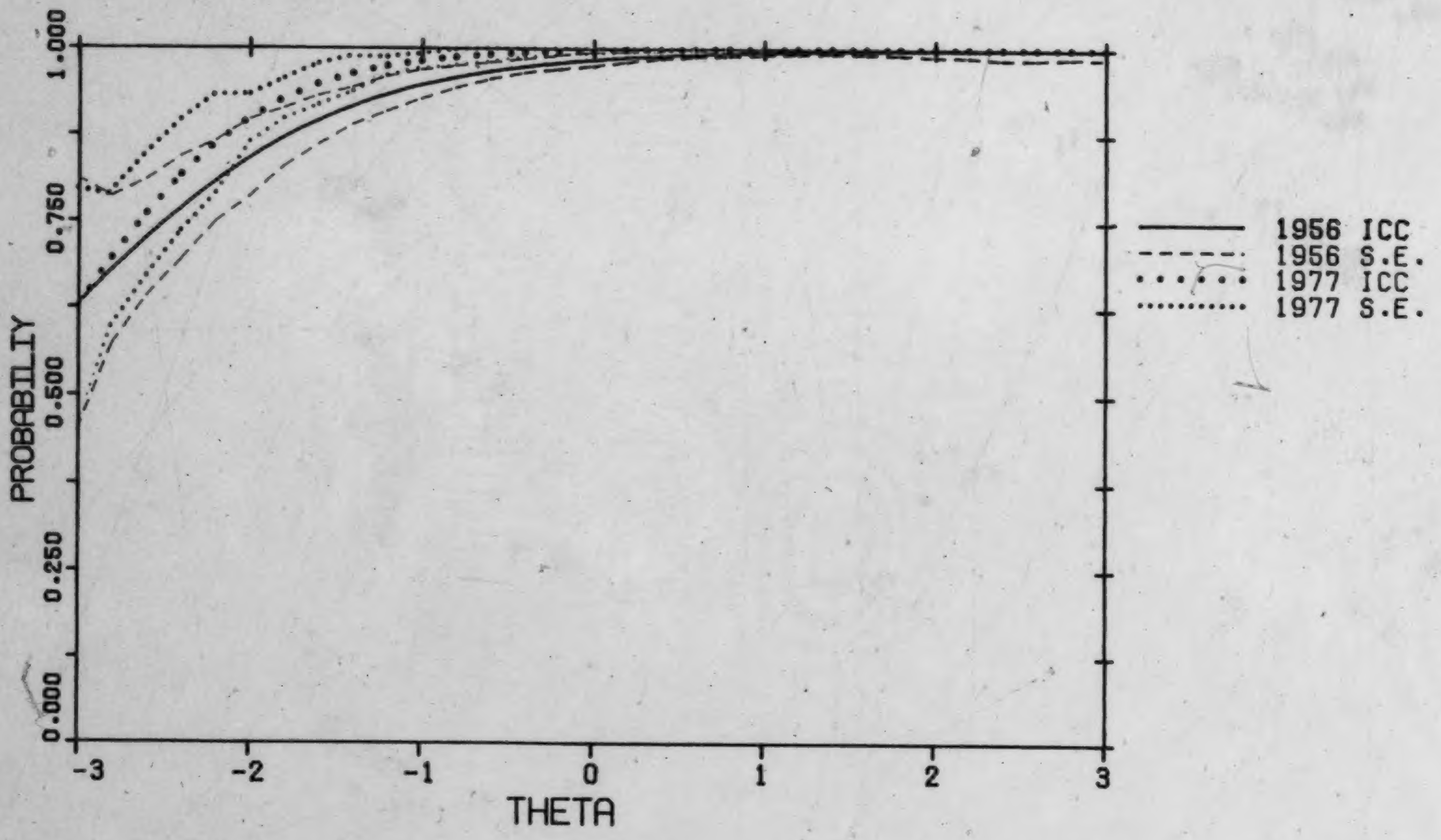


Figure 104

ICC'S FOR ITEM NUMBER 17 : CMM

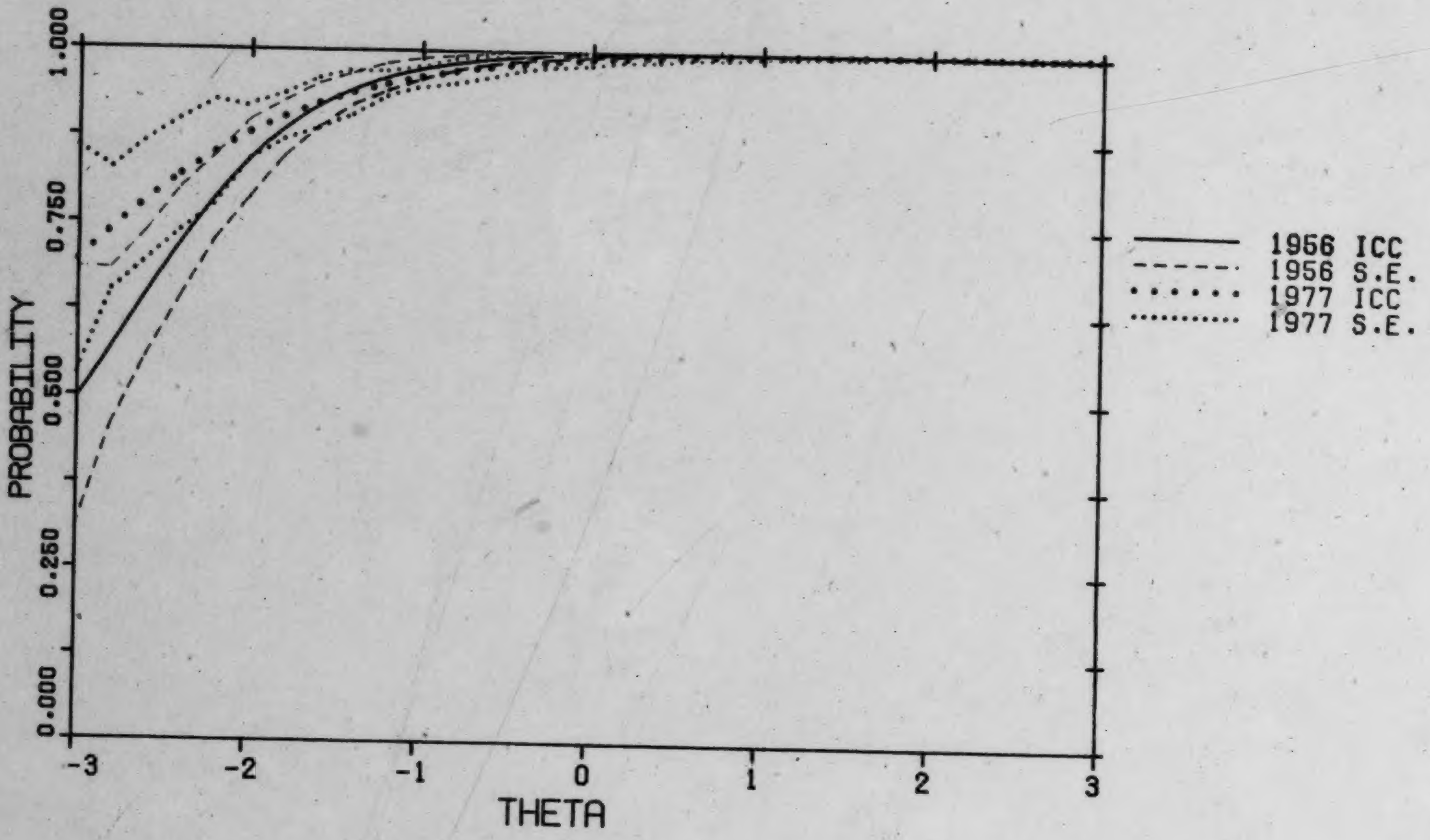


Figure 105

ICC'S FOR ITEM NUMBER 18 : CMM

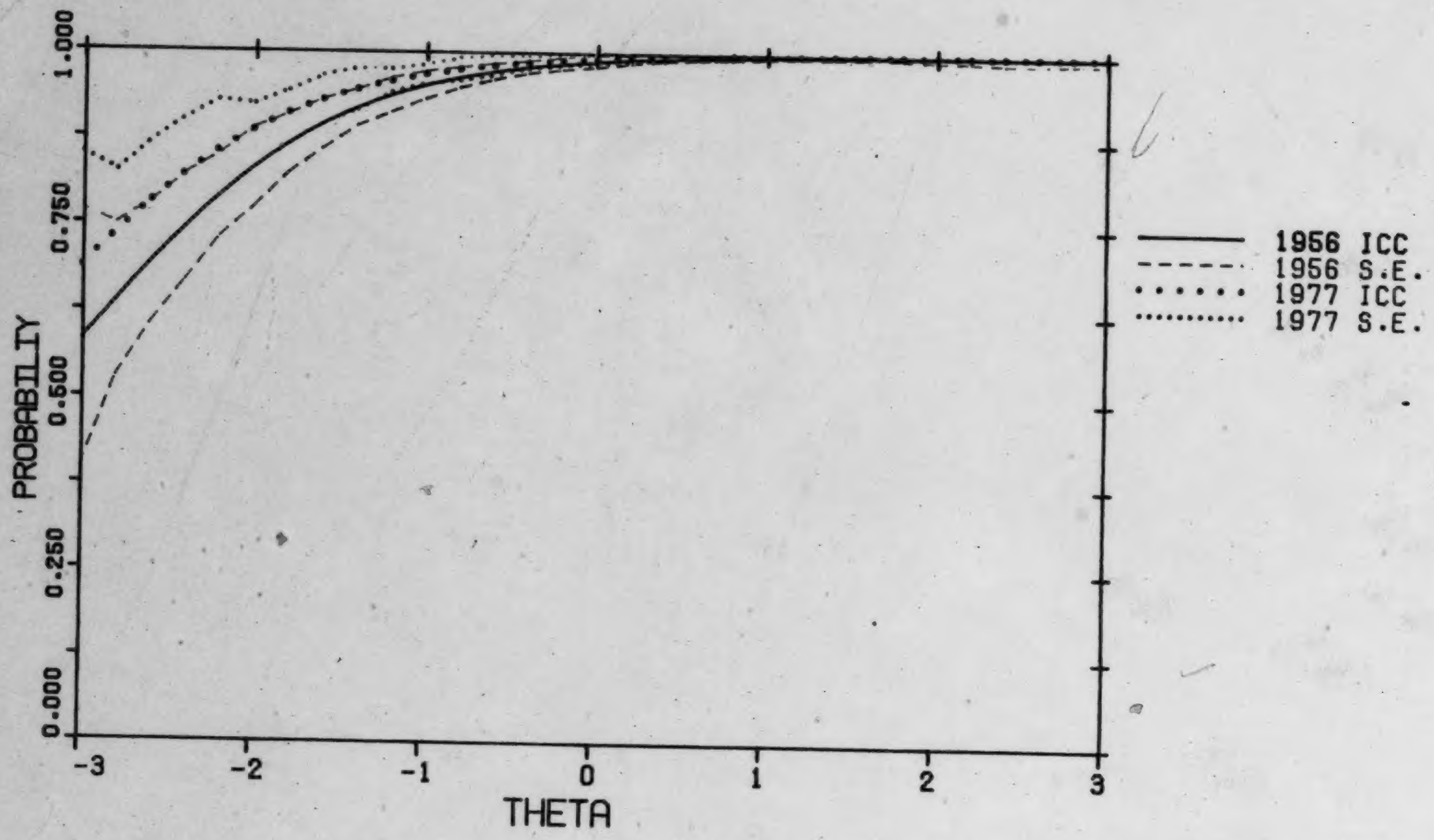


Figure 106

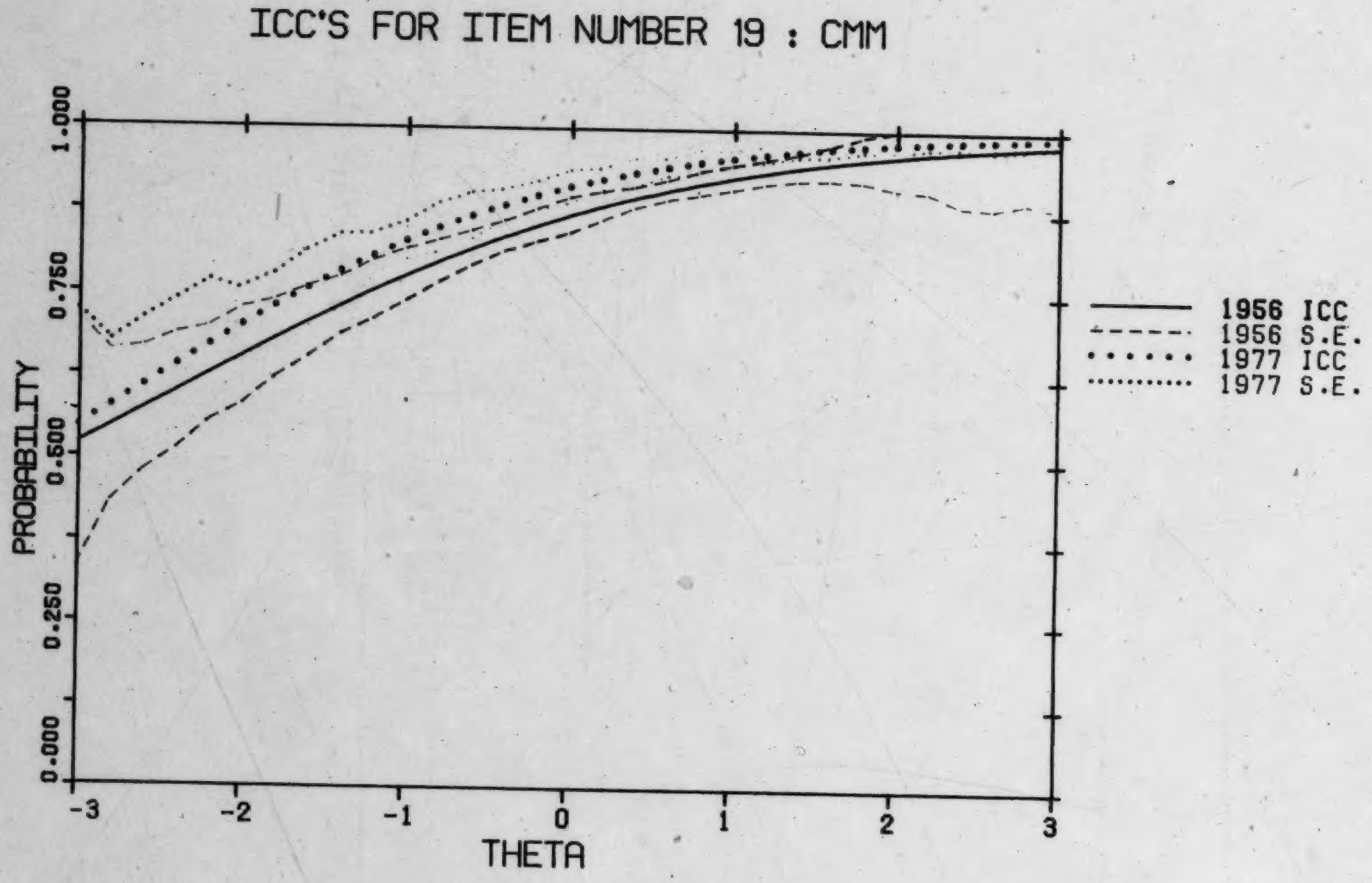

ICC'S FOR ITEM NUMBER 19 : CMM
PROBABILITY
1.000
0.750
0.500
0.250
0.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 107

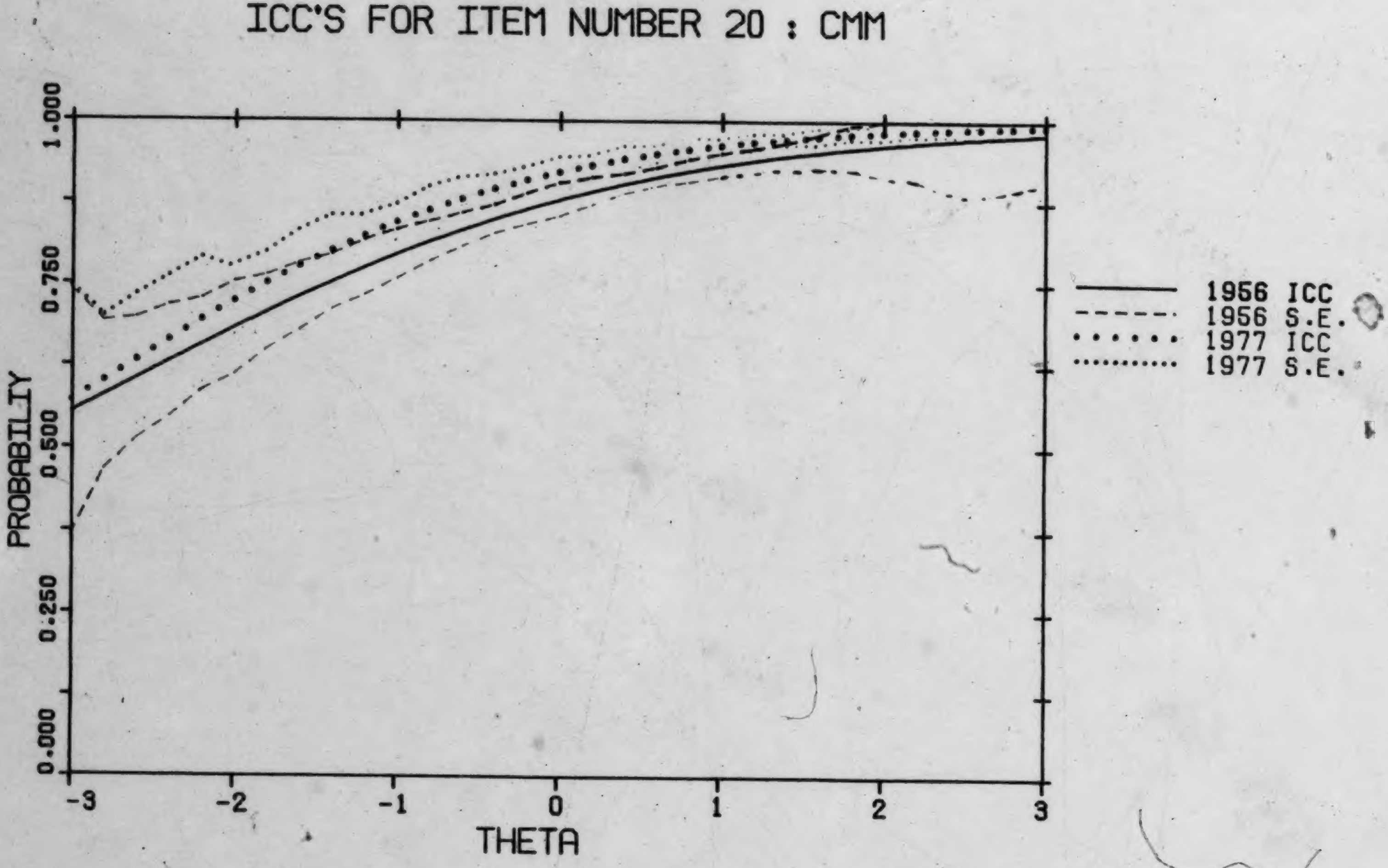

ICC'S FOR ITEM NUMBER 20 : CMM
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 108

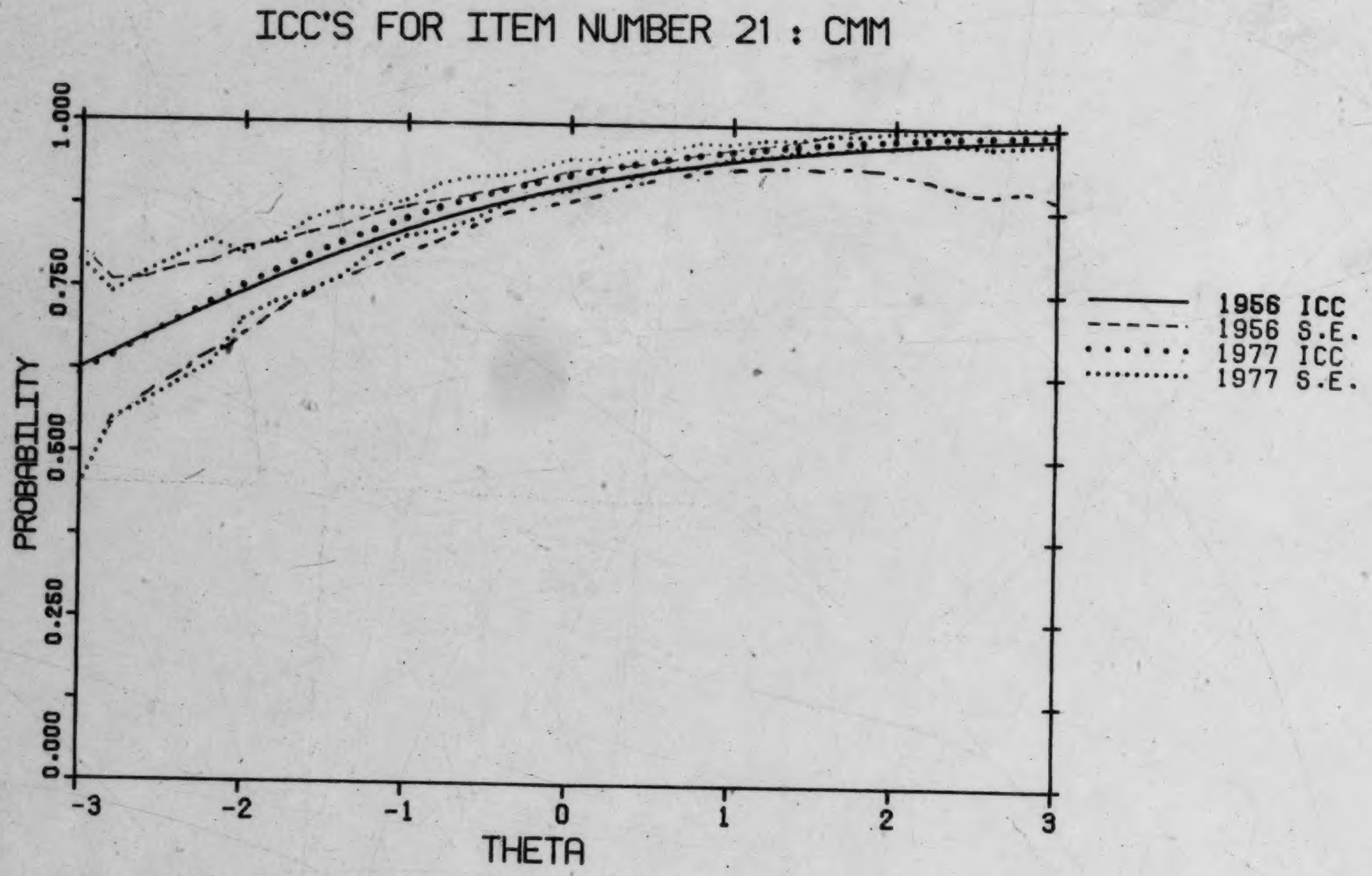

ICC'S FOR ITEM NUMBER 21 : CMM
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 109

ICC'S FOR ITEM NUMBER 22 : CMM

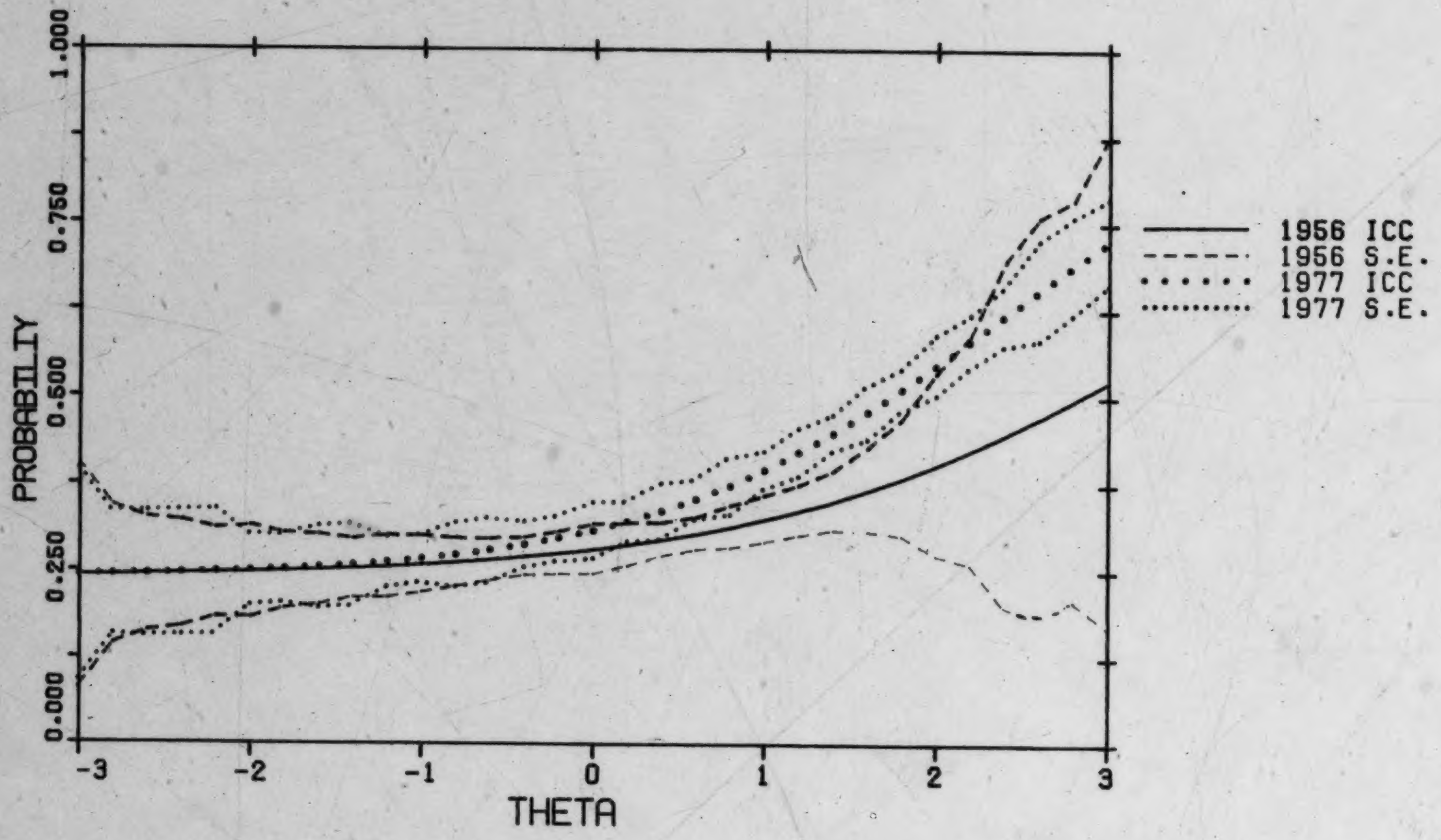


Figure 110

ICC'S FOR ITEM NUMBER 23 : CMM

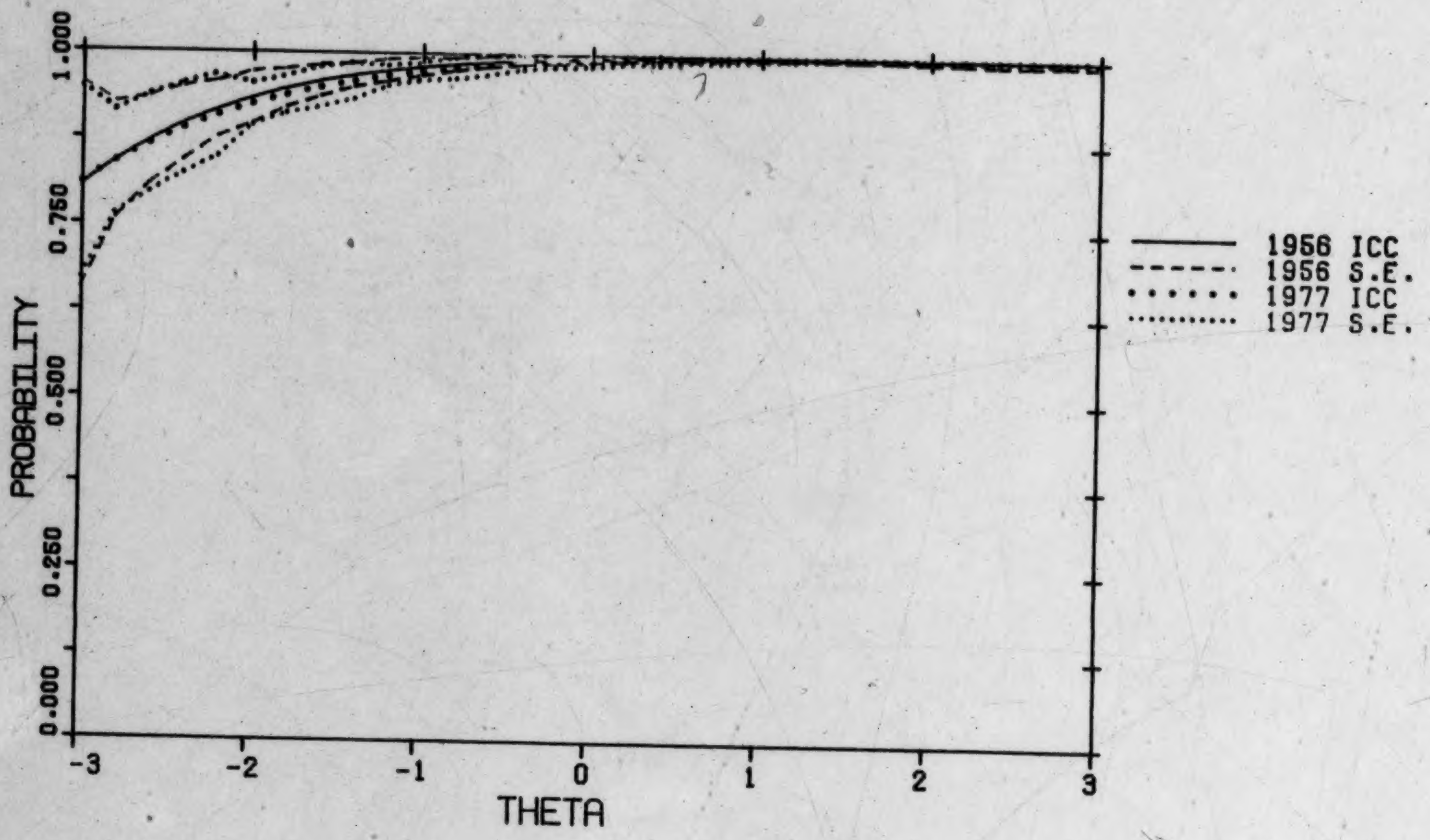


Figure 111

ICC'S FOR ITEM NUMBER 24 : CMM

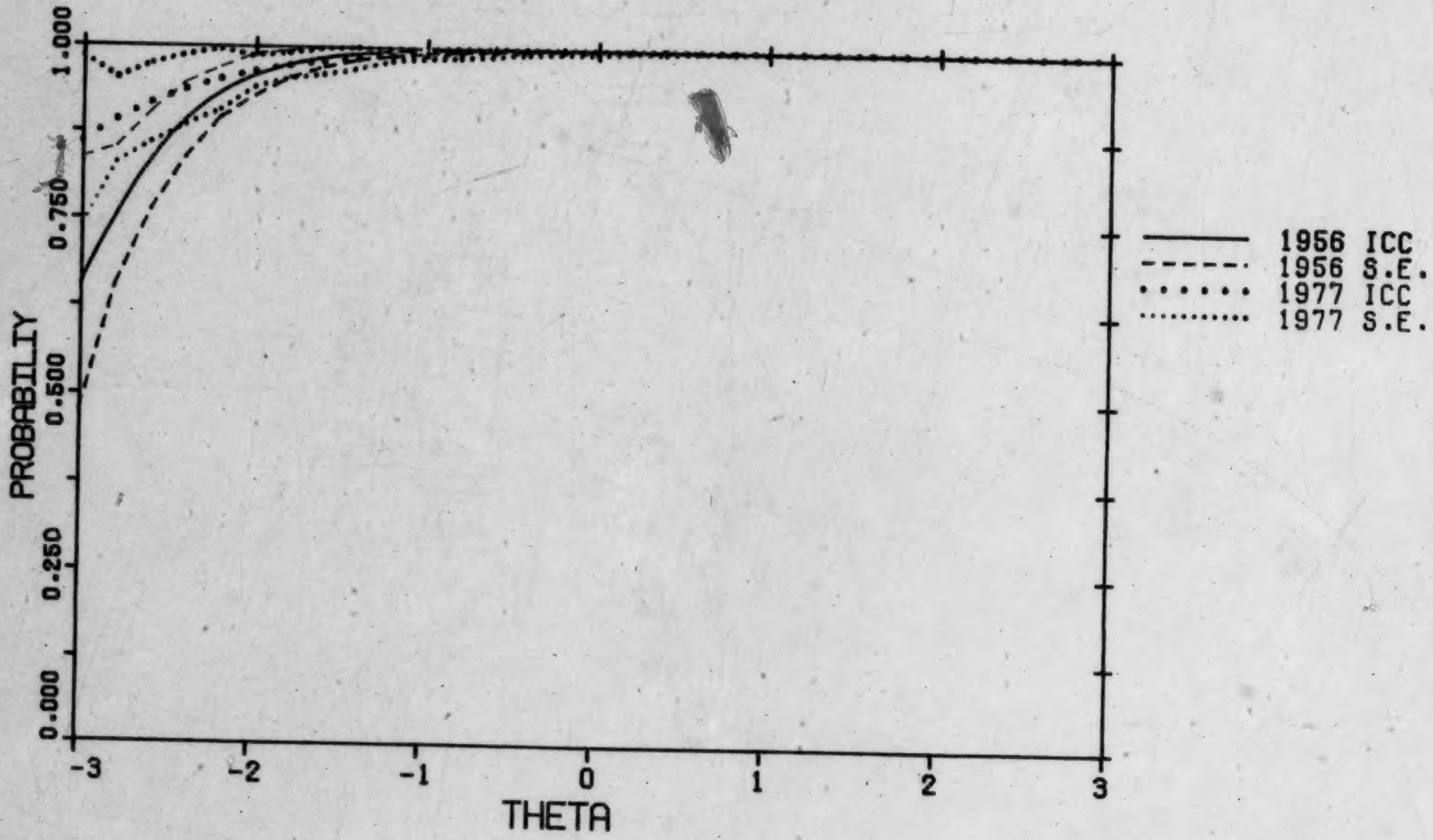


Figure 112

ICC'S FOR ITEM NUMBER 25 : CMM

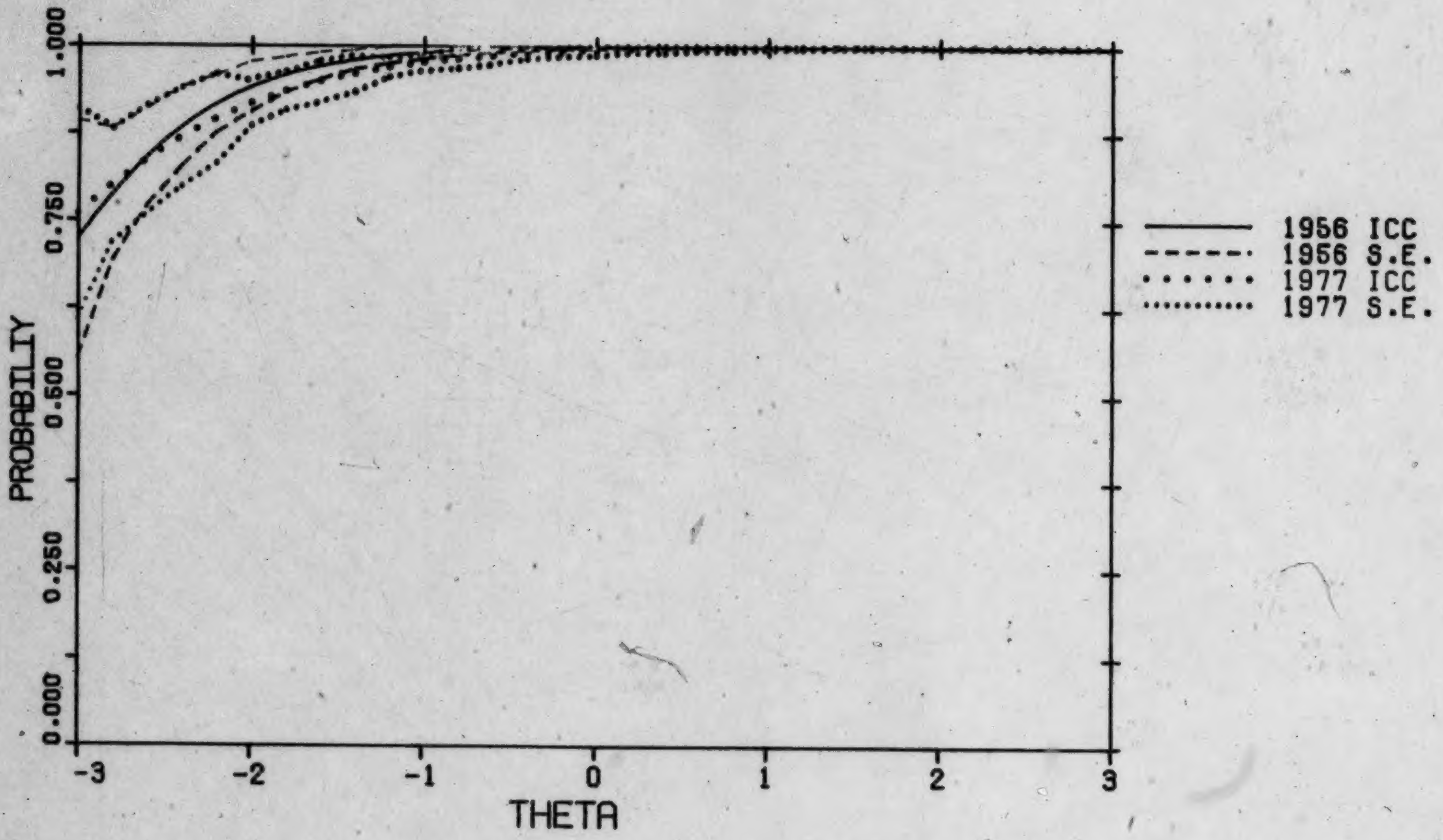


Figure 113

ICC'S FOR ITEM NUMBER 27 : CMM

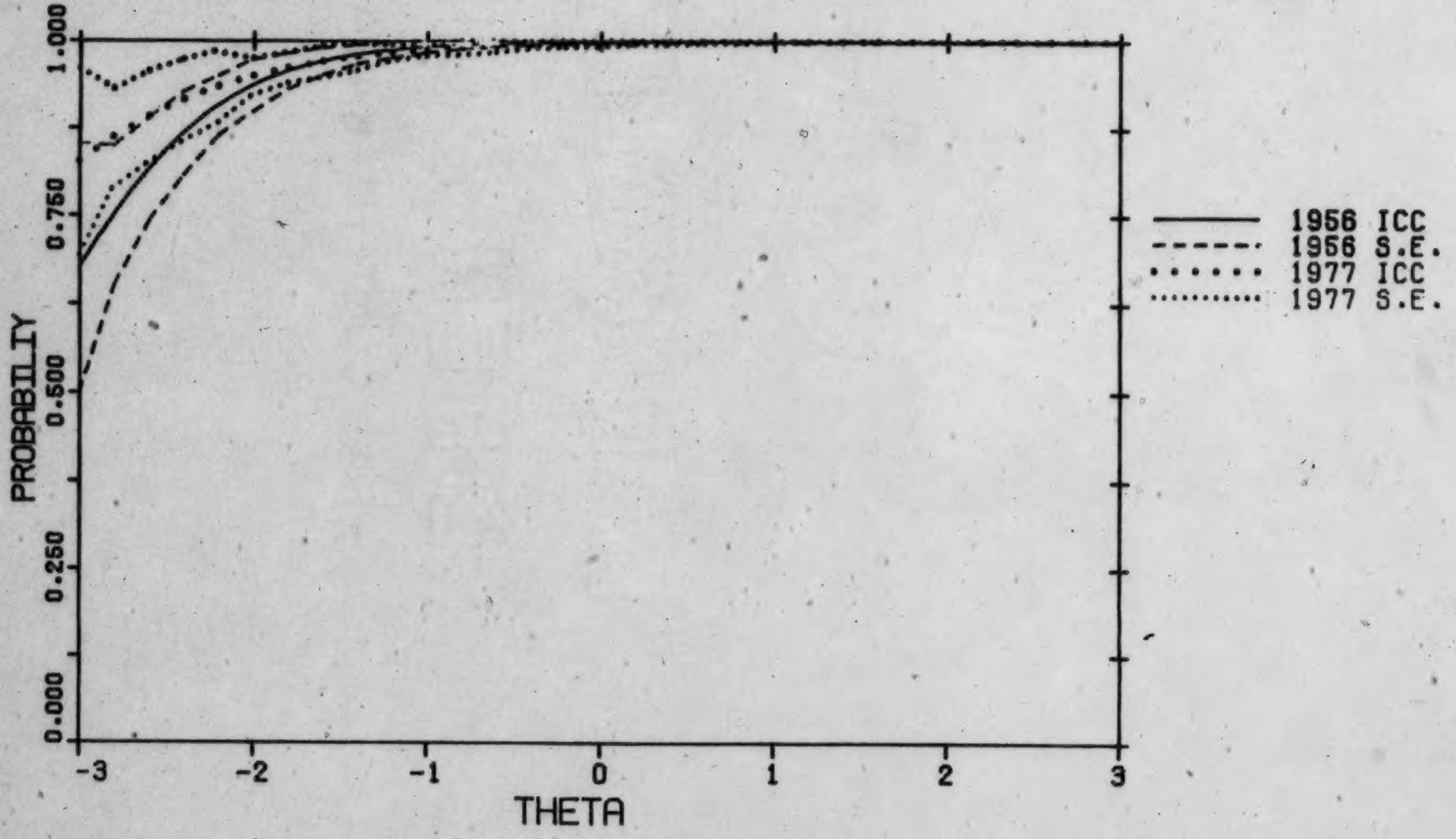


Figure 114

ICC'S FOR ITEM NUMBER 28 : CMM

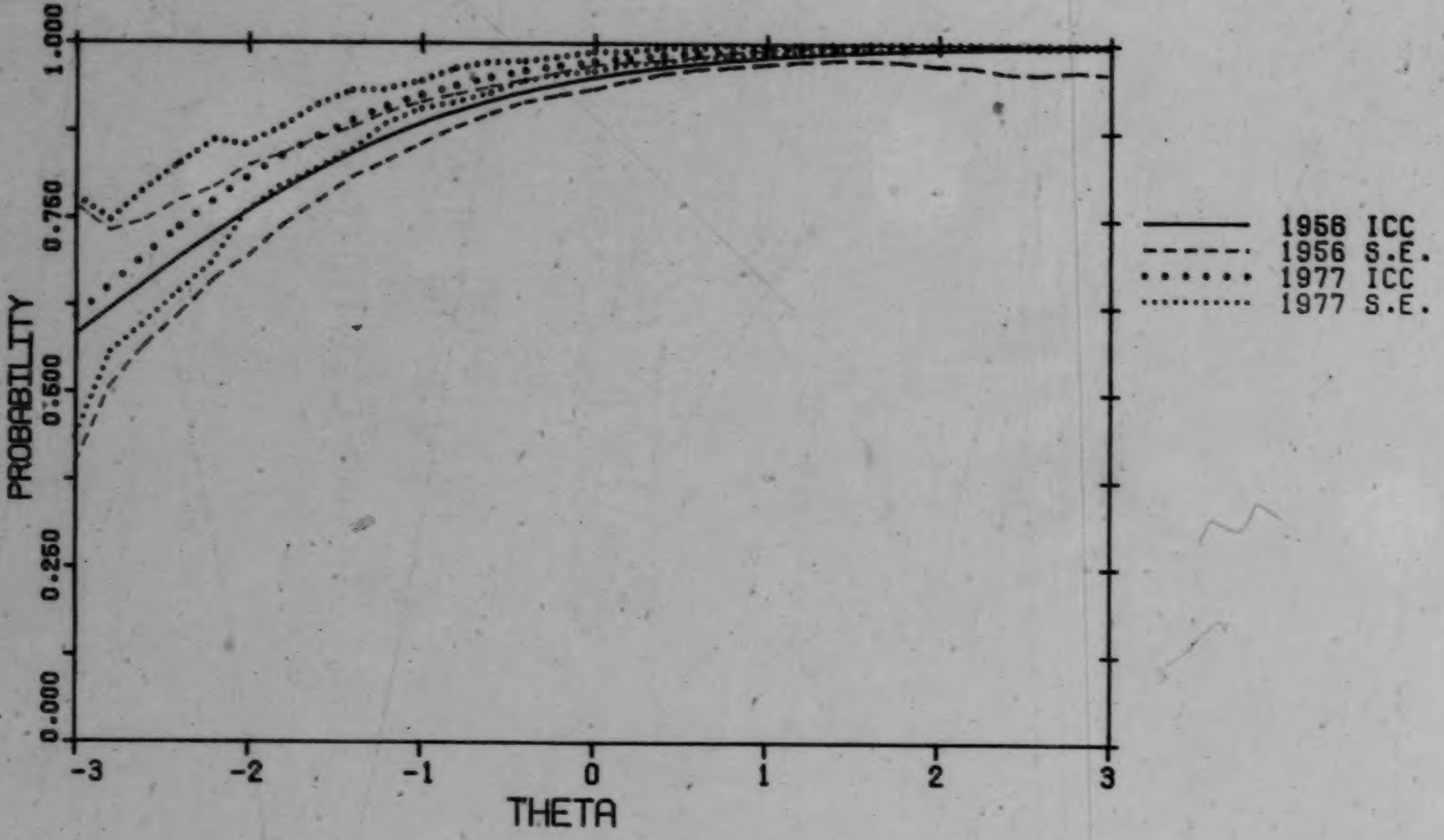


Figure 115

ICC'S FOR ITEM NUMBER 29 : CMM

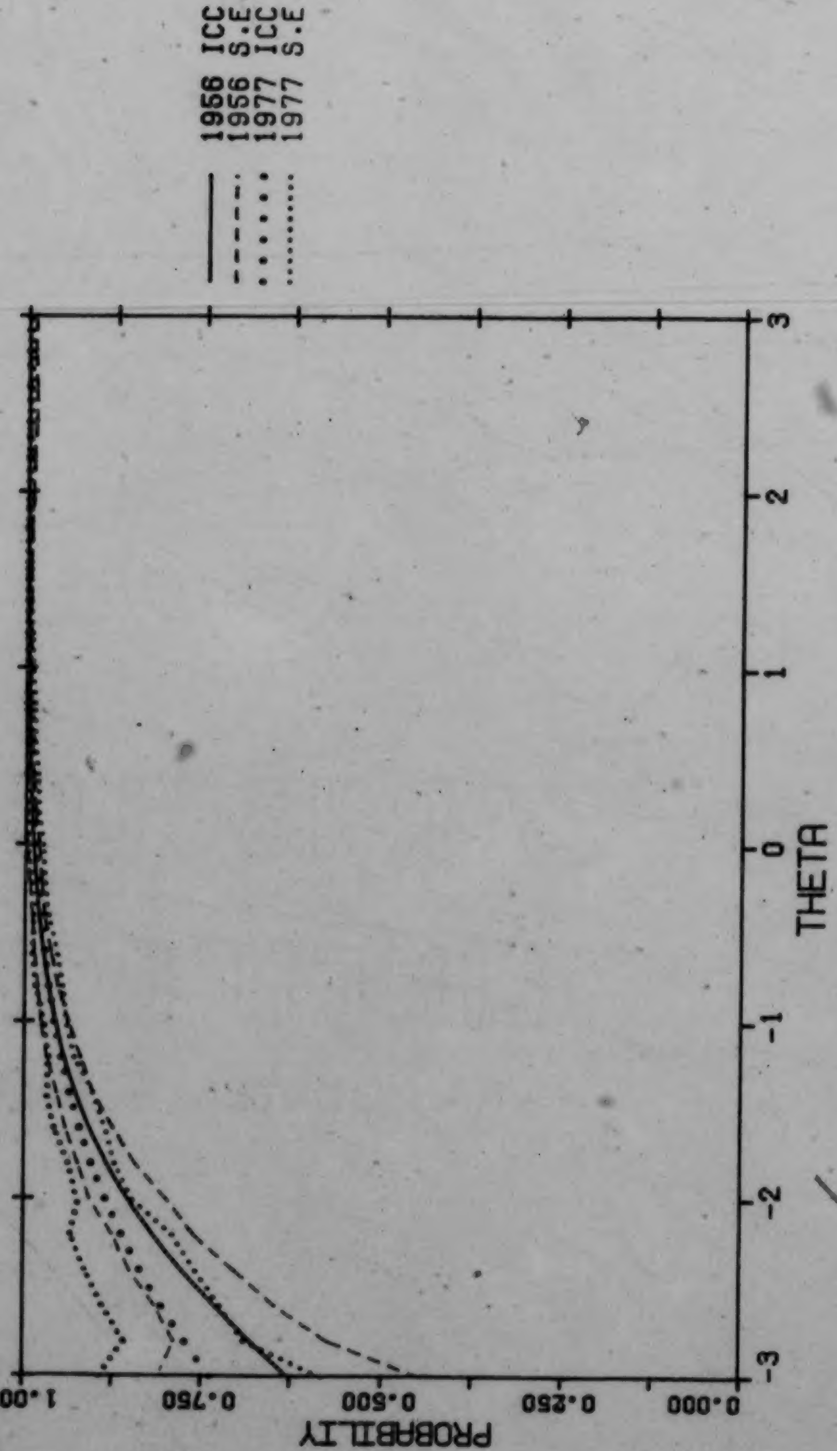


Figure 116

ICC'S FOR ITEM NUMBER 30 : CMM

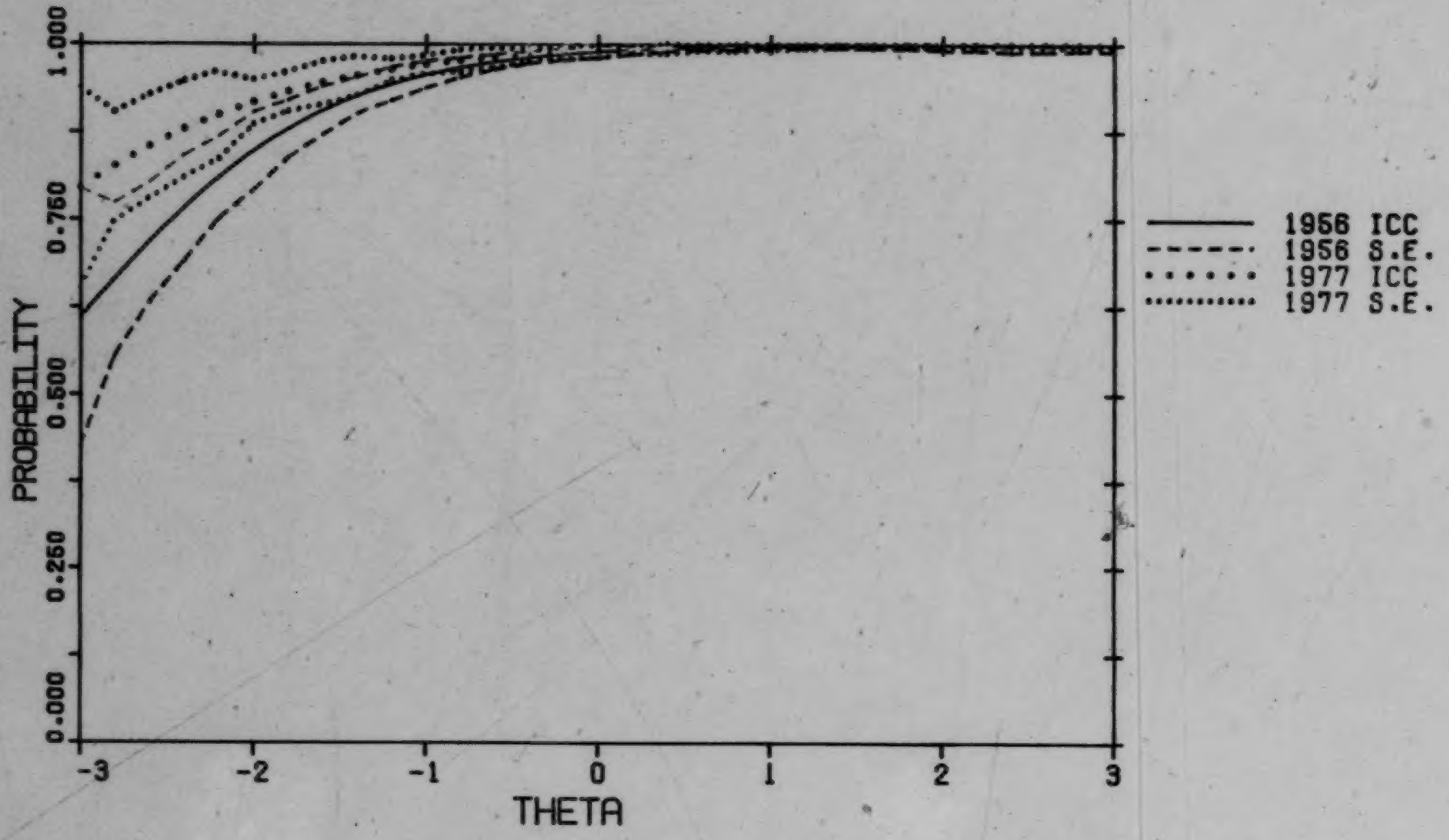


Figure 117

ICC'S FOR ITEM NUMBER 31 : CMM

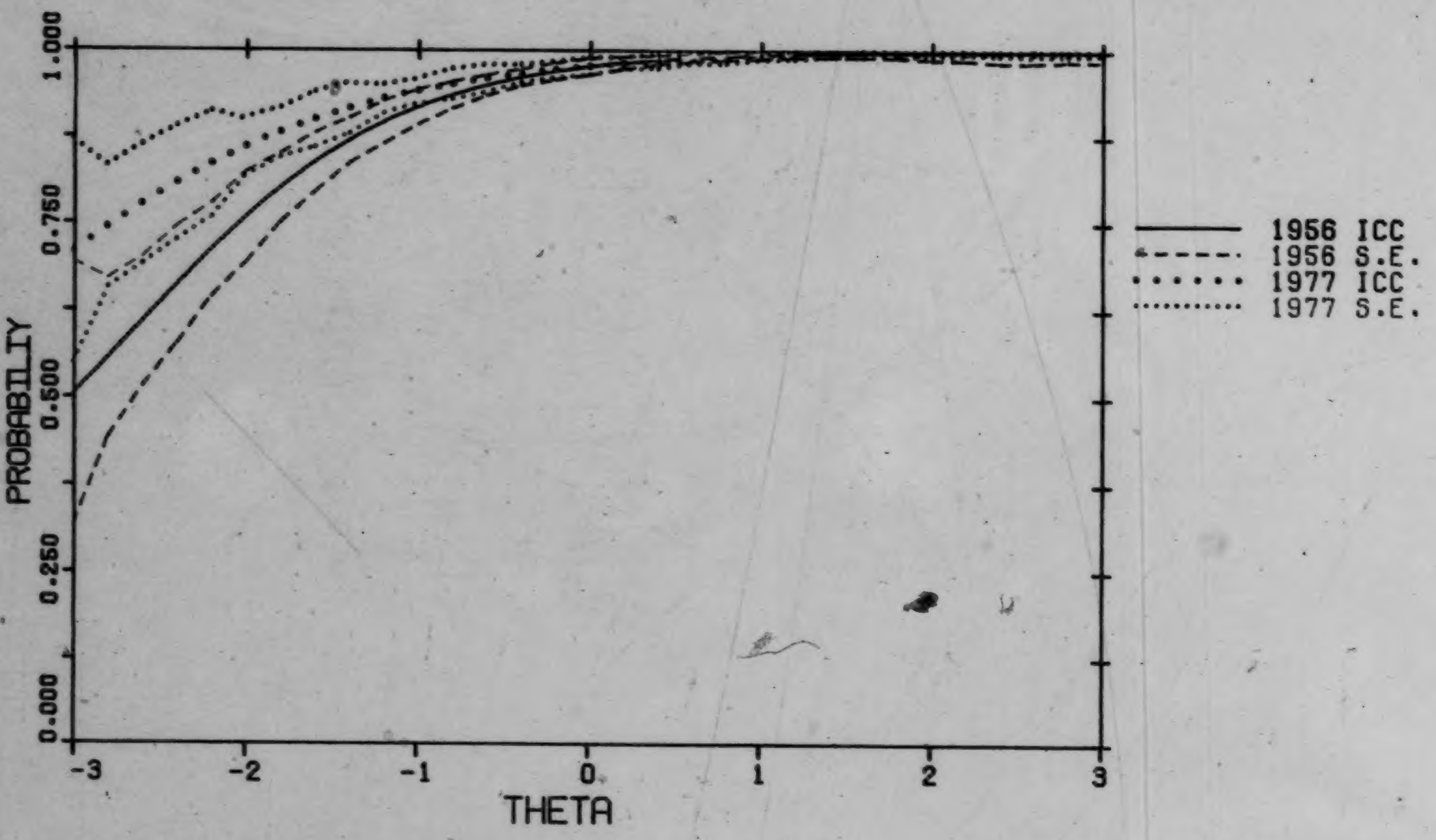


Figure 118

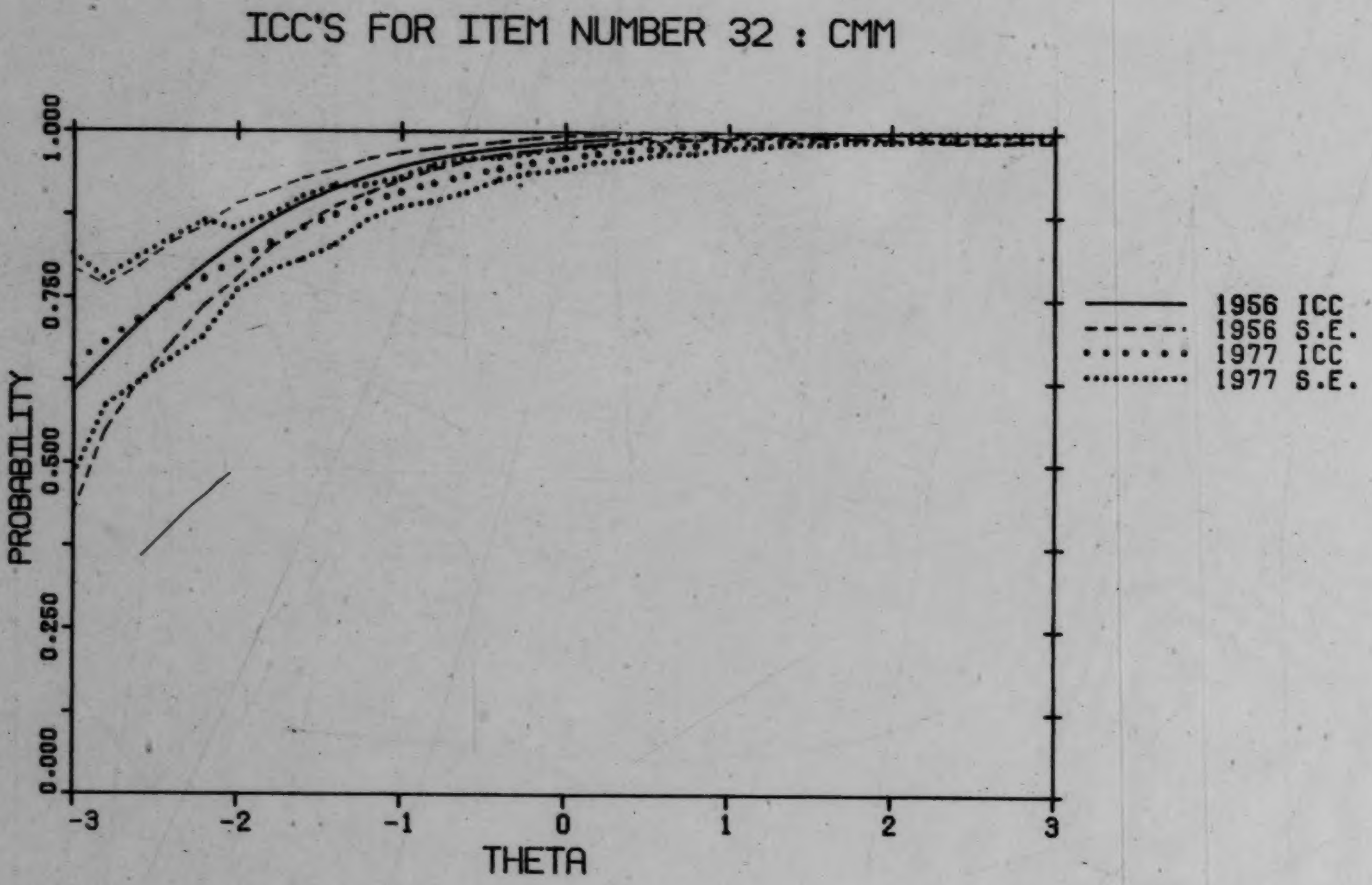

ICC'S FOR ITEM NUMBER 32 : CMM
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 119

ICC'S FOR ITEM NUMBER 33 : CMM

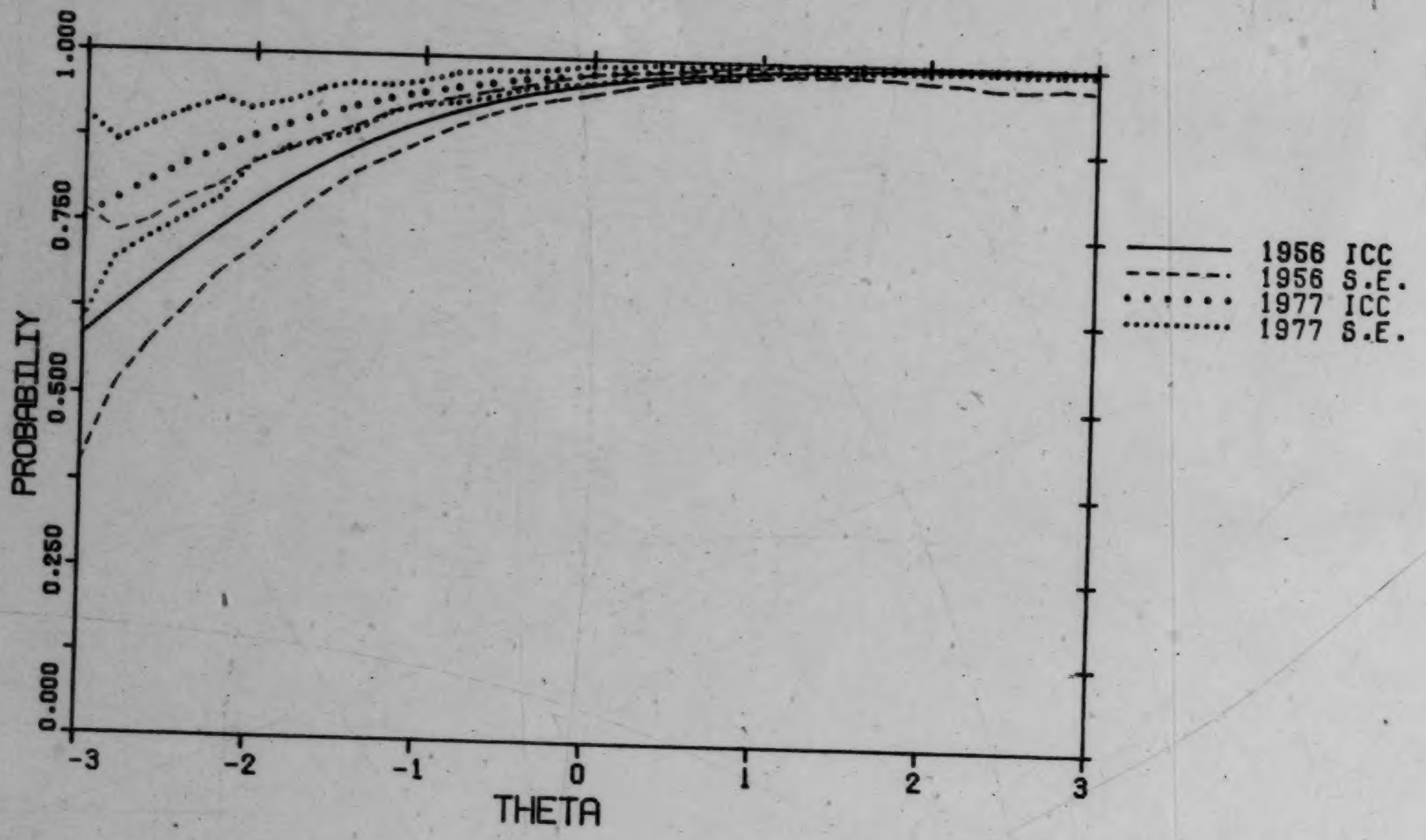


Figure 120

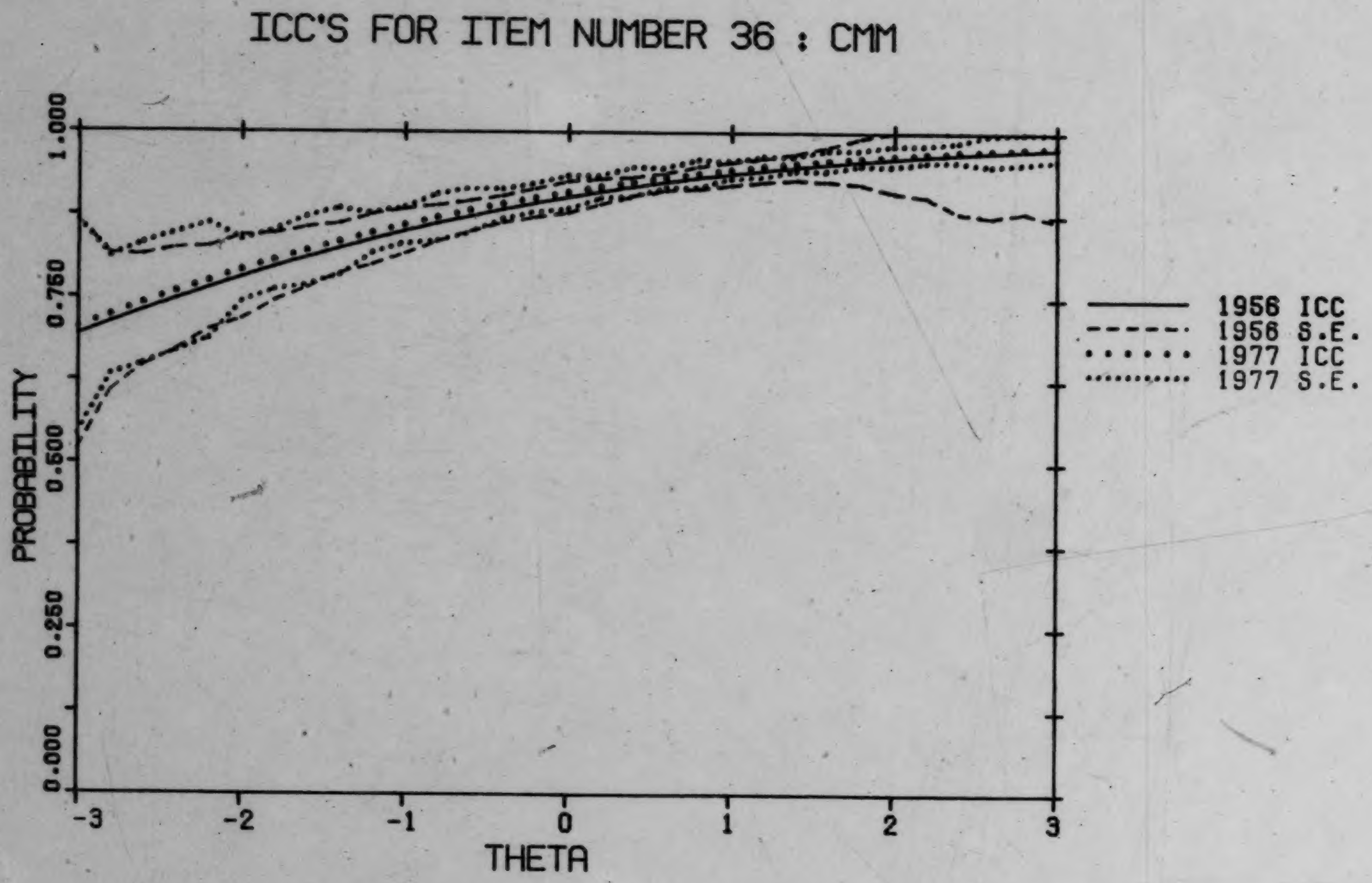

ICC'S FOR ITEM NUMBER 36 : CMM
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 121

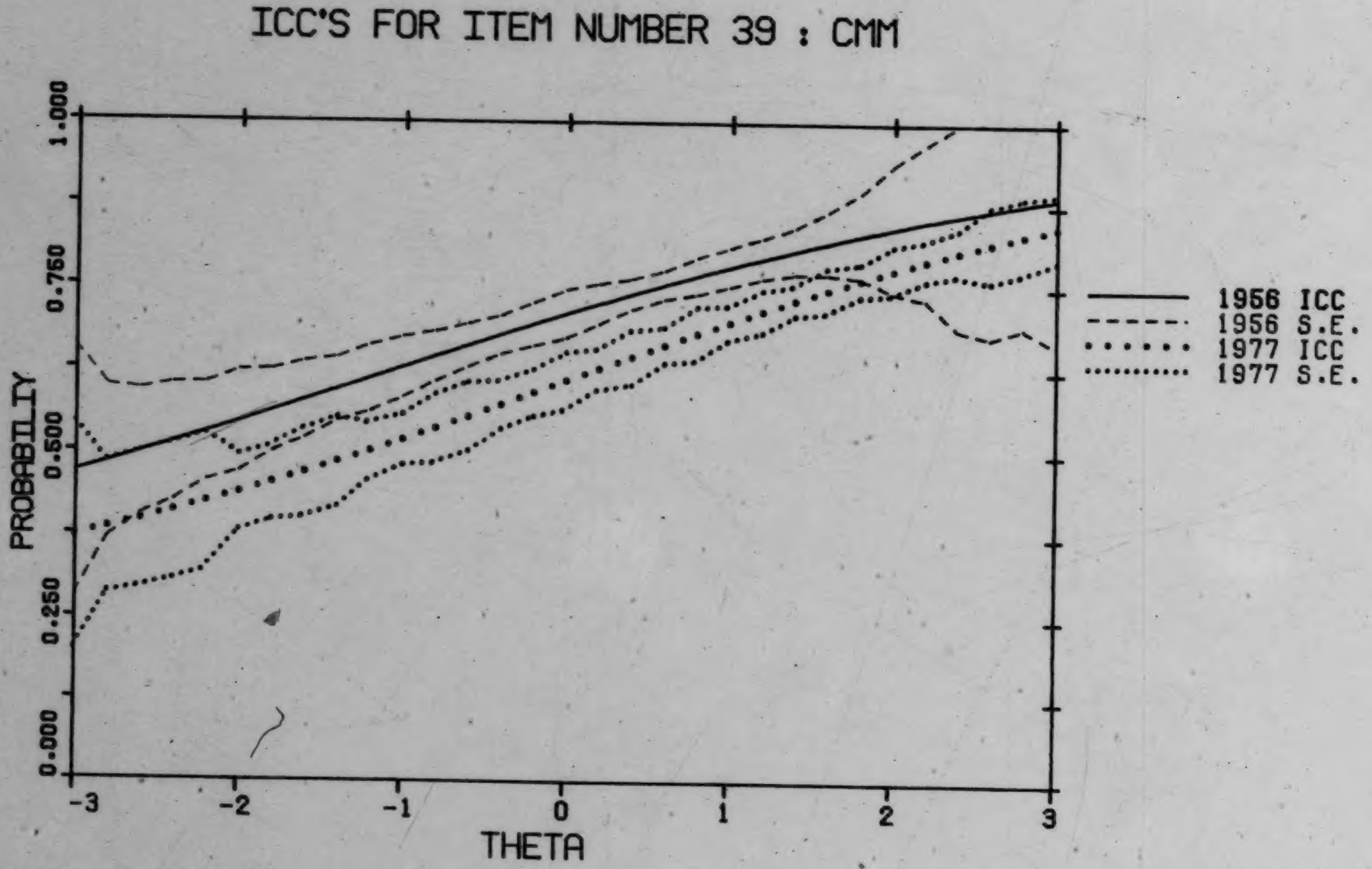

ICC'S FOR ITEM NUMBER 39 : CMM
PROBABILITY
1.000
0.750
0.500
0.250
0.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 122

ICC'S FOR ITEM NUMBER 40 : CMM

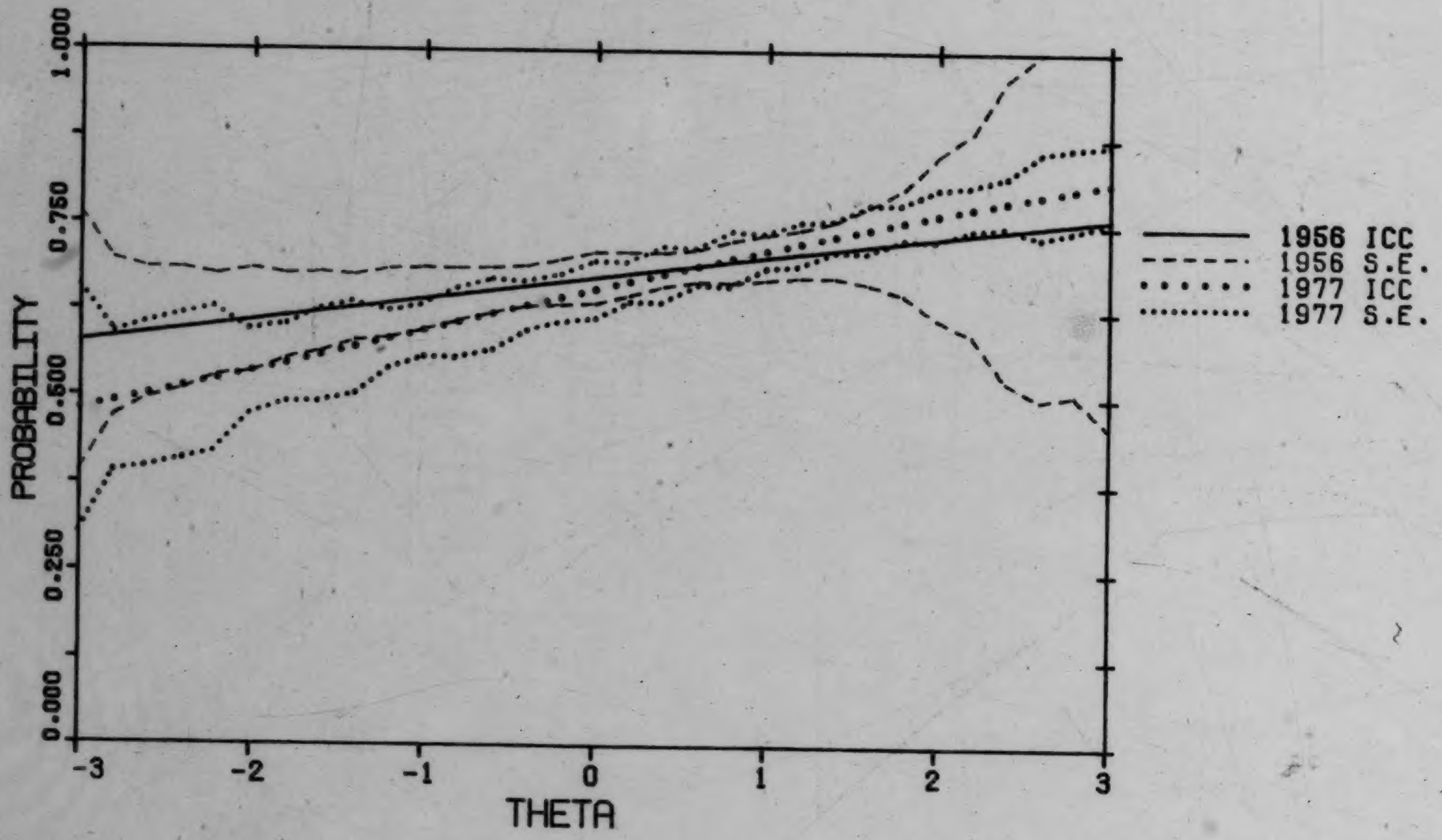


Figure 123

ICC'S FOR ITEM NUMBER 41 : CMM

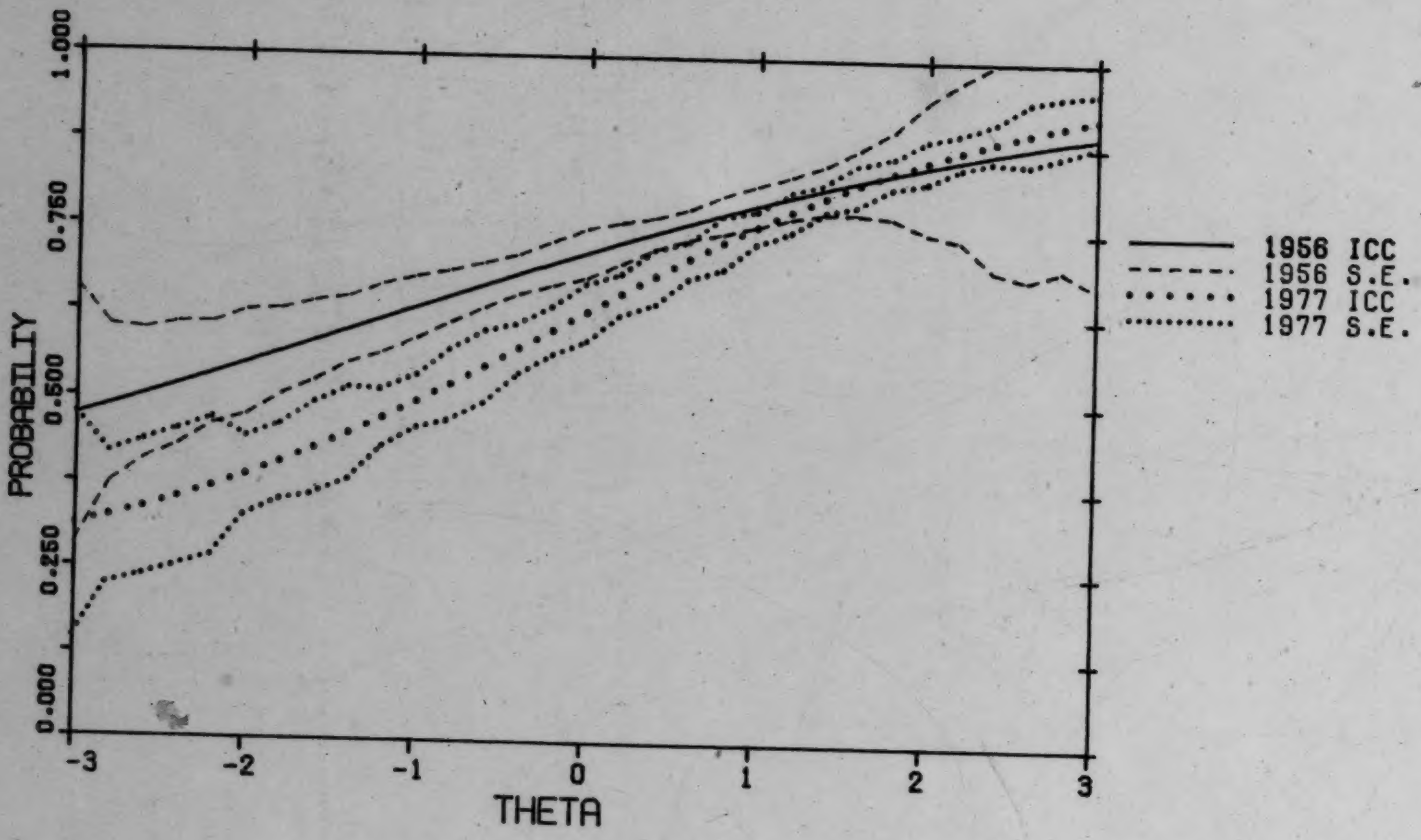


Figure 124

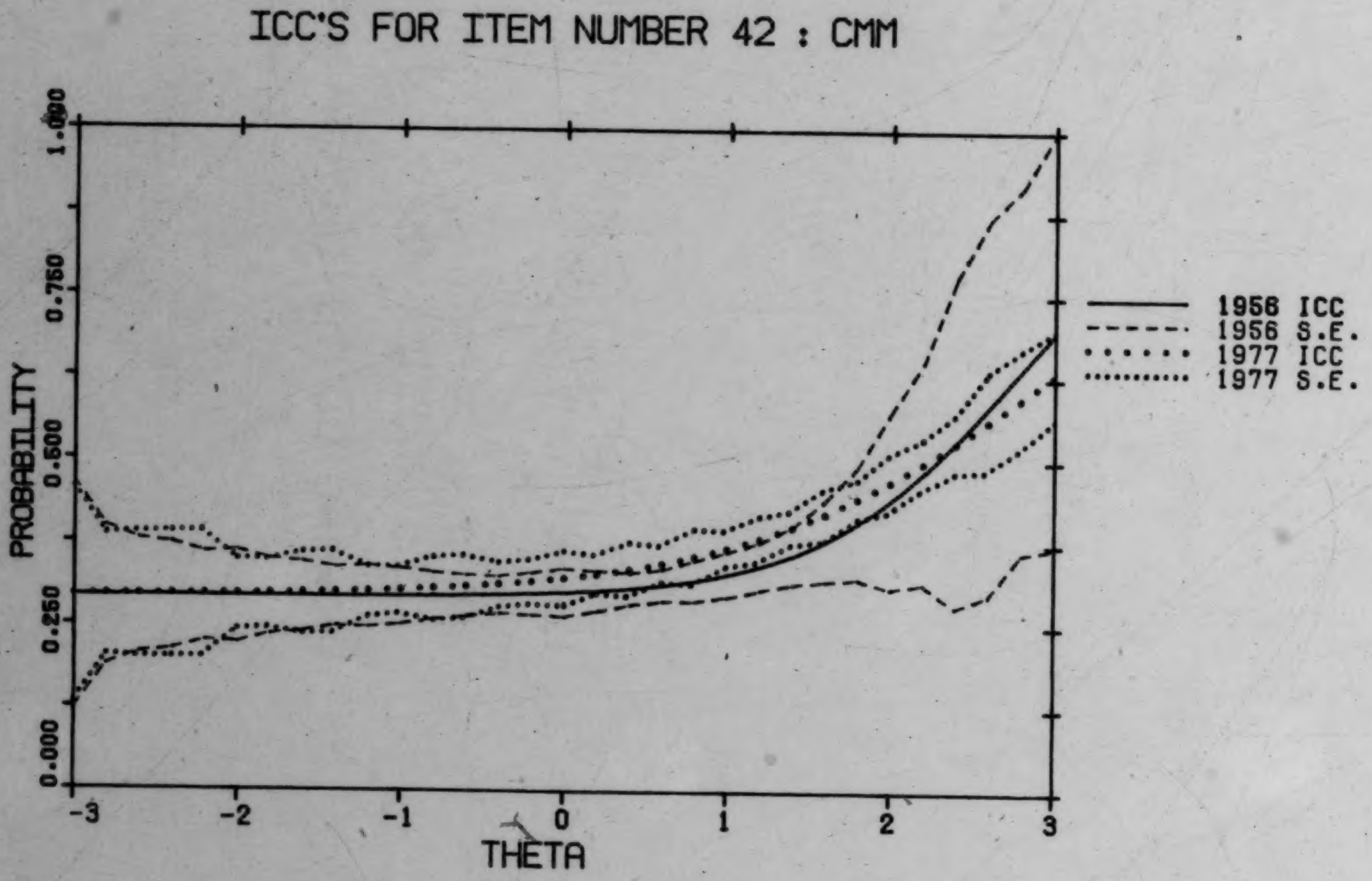

ICC'S FOR ITEM NUMBER 42 : CMM
PROBABILITY
0.000
0.250
0.500
0.750
1.000
-3
-2
-1
0
1
2
3
THETA
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 125

ICC'S FOR ITEM NUMBER 45 : CMM

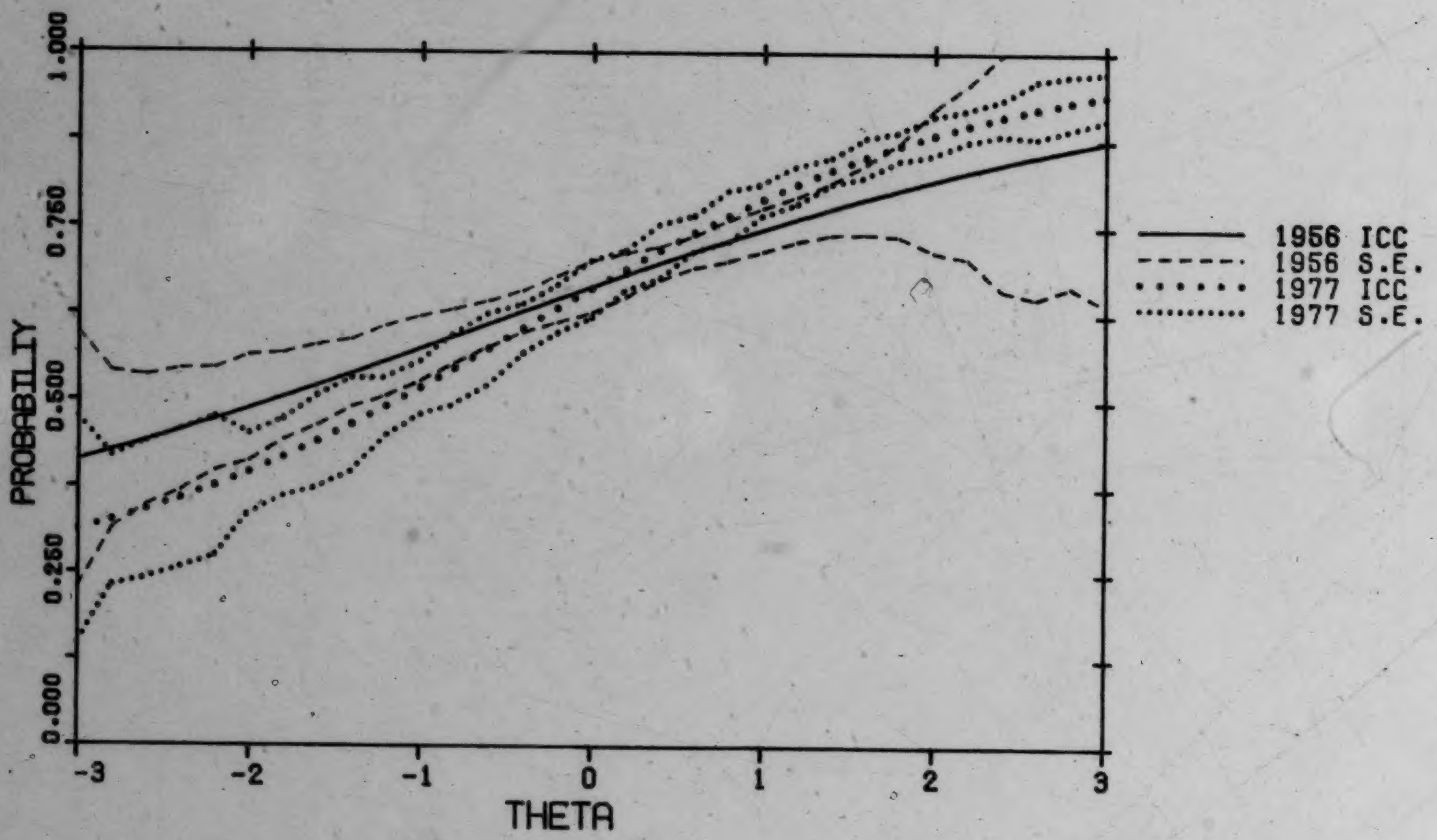


Figure 126

ICC'S FOR ITEM NUMBER 46 : CMM

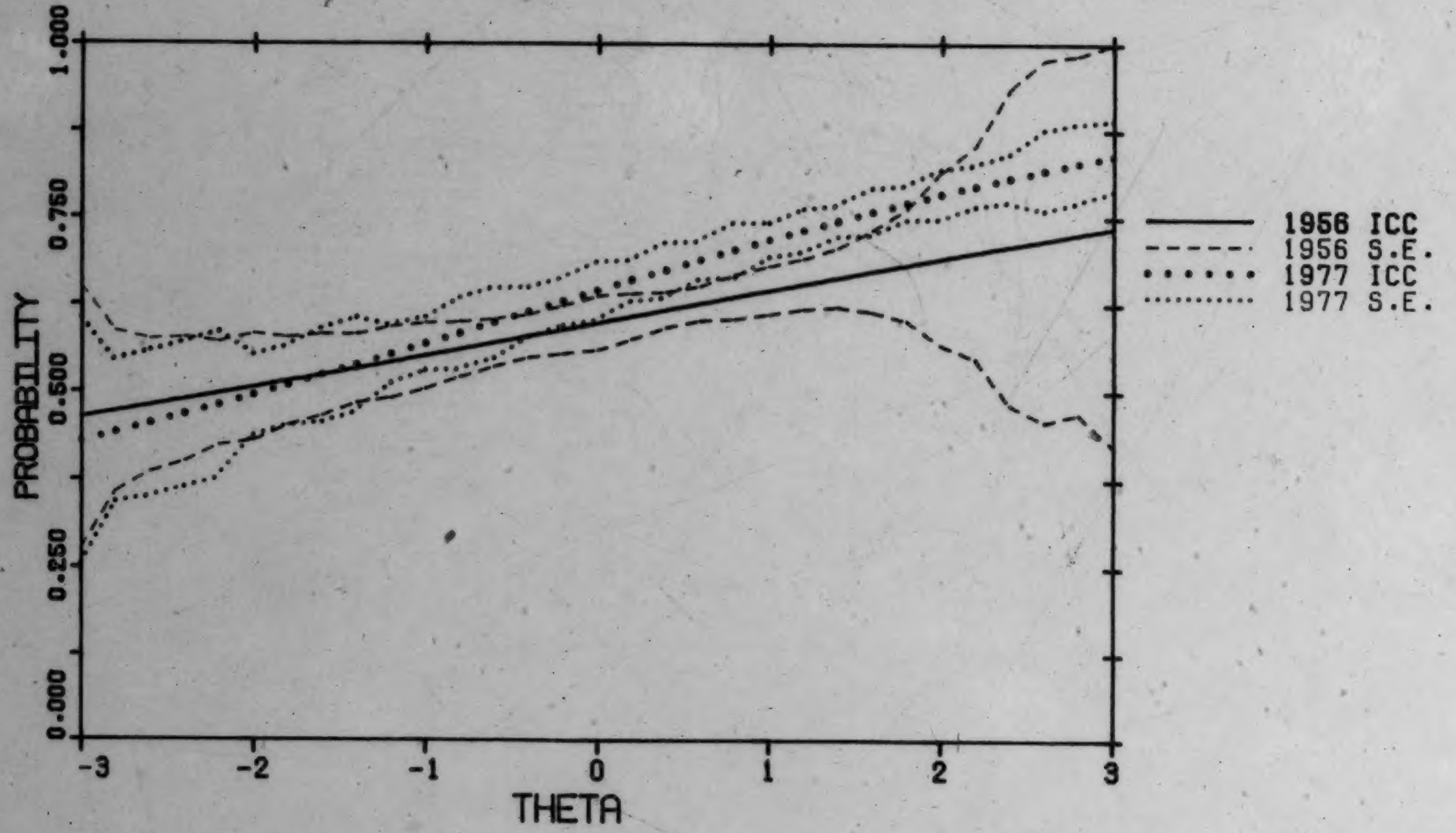


Figure 127

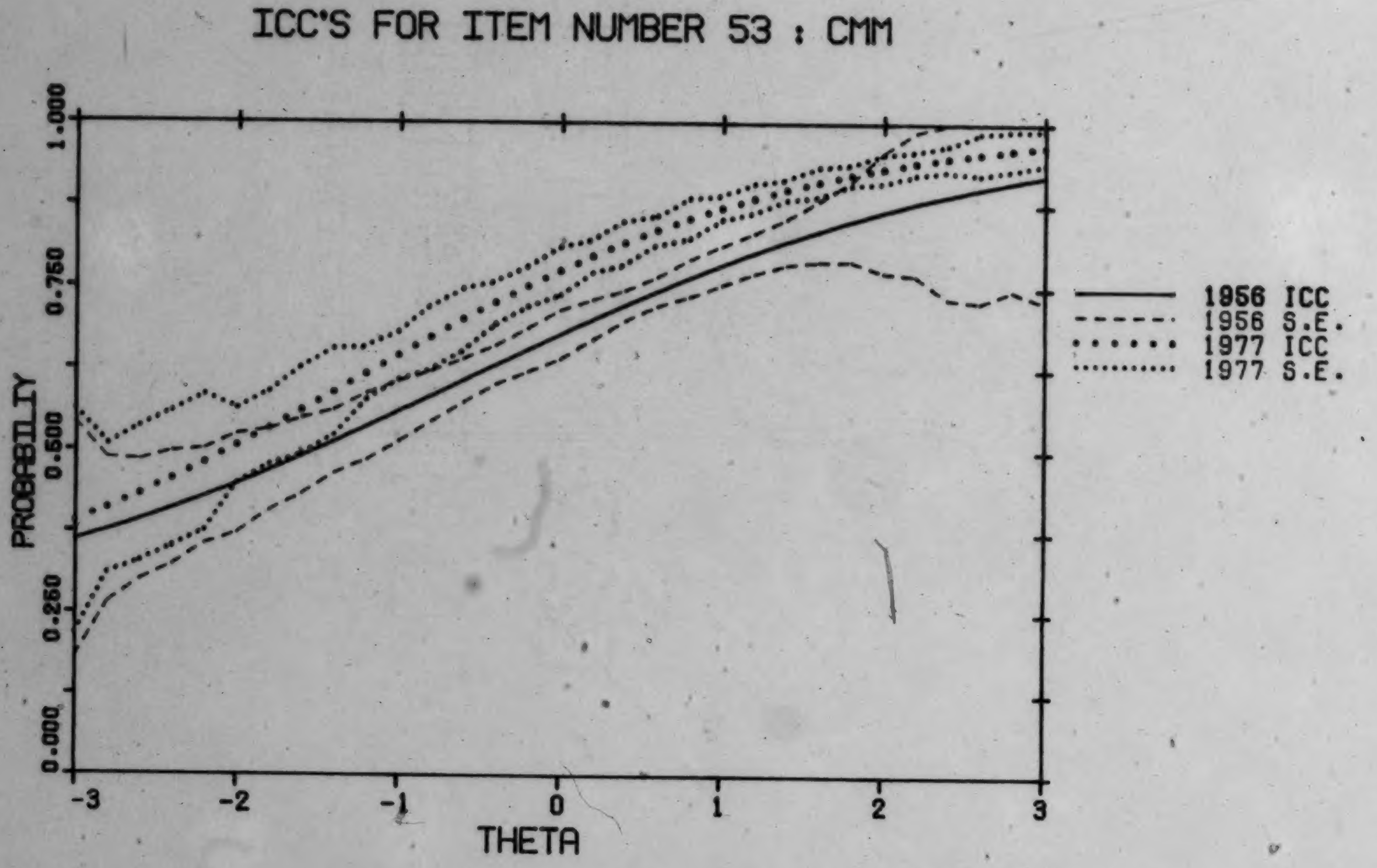

ICC'S FOR ITEM NUMBER 53 : CMM
PROBABILITY
0.000
0.250
0.500
0.750
1.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 128

ICC'S FOR ITEM NUMBER 54 : CMM

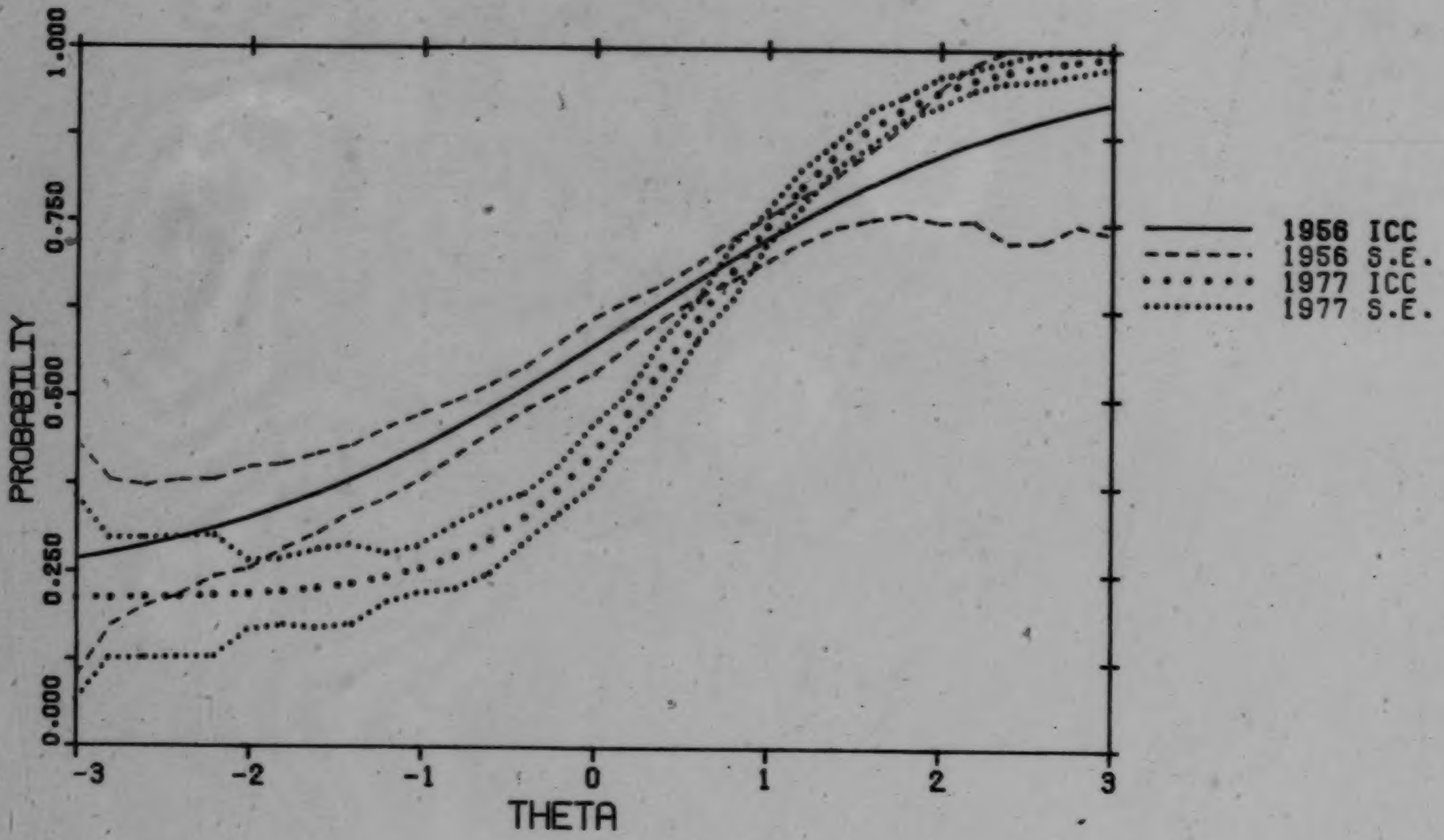


Figure 129

ICC'S FOR ITEM NUMBER 55 : CMM

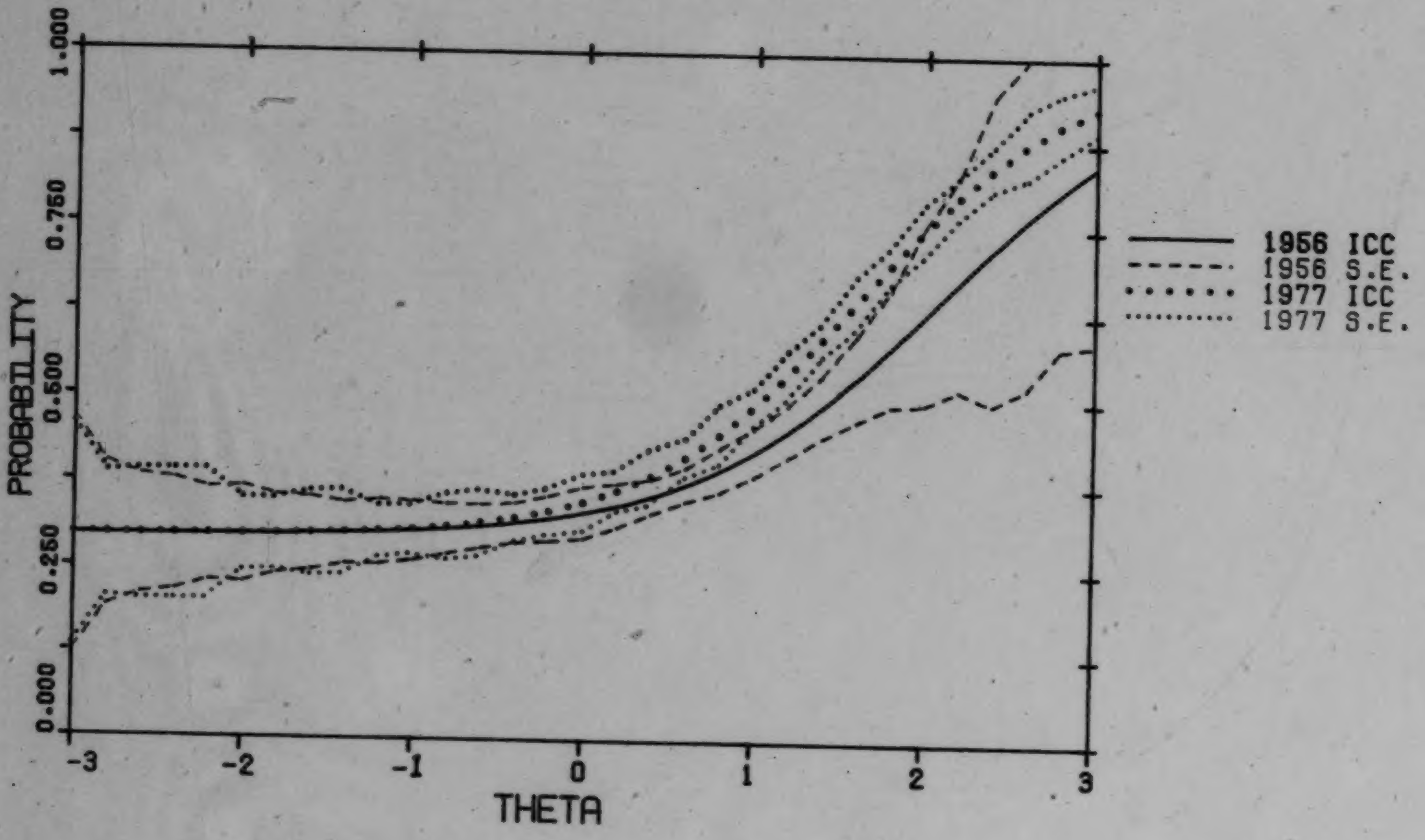


Figure 130

ICC'S FOR ITEM NUMBER 56 : CMM

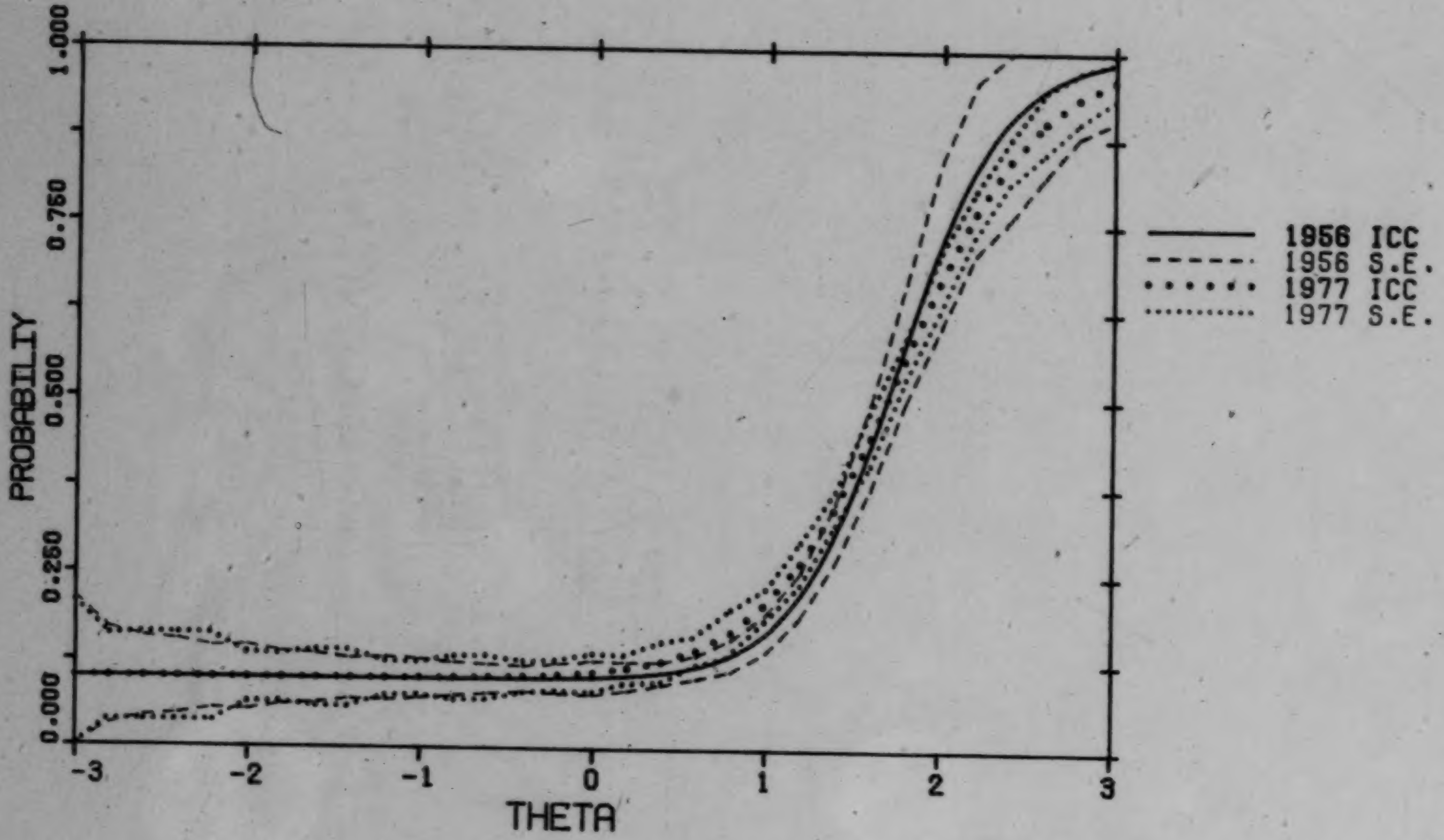


Figure 131

ICC'S FOR ITEM NUMBER 57 : CMM

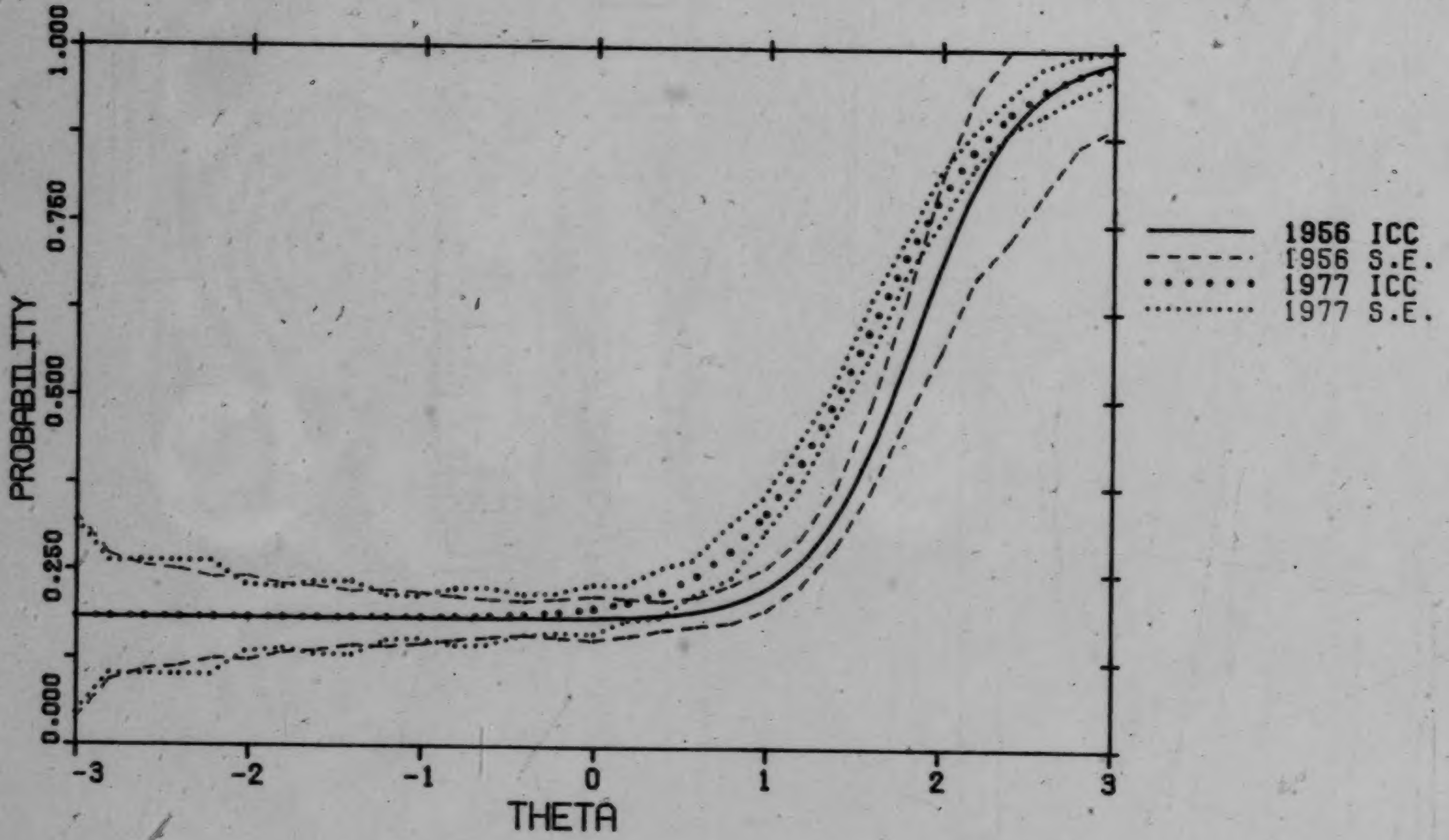


Figure 132

ICC'S FOR ITEM NUMBER 59 : CMM

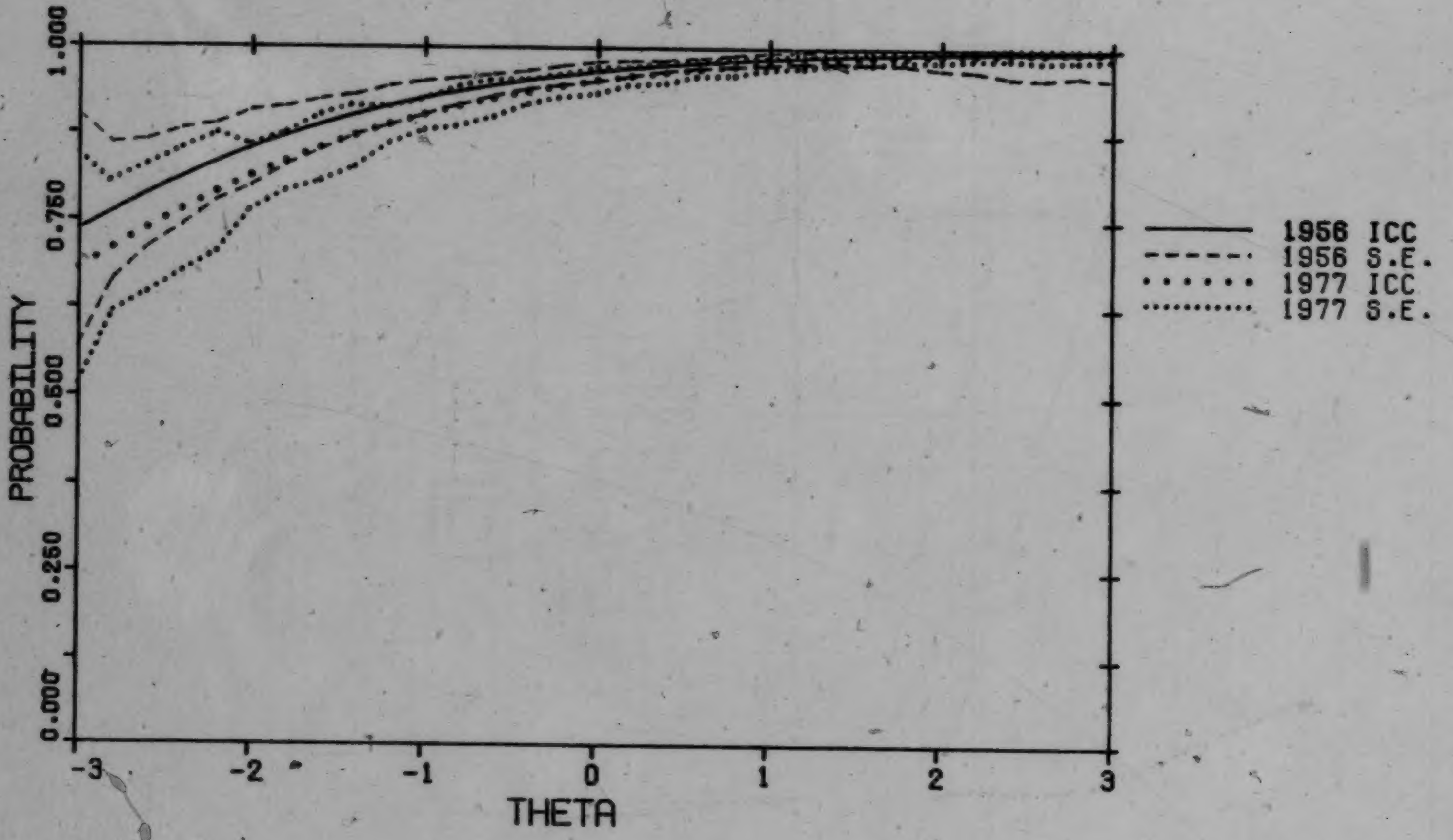
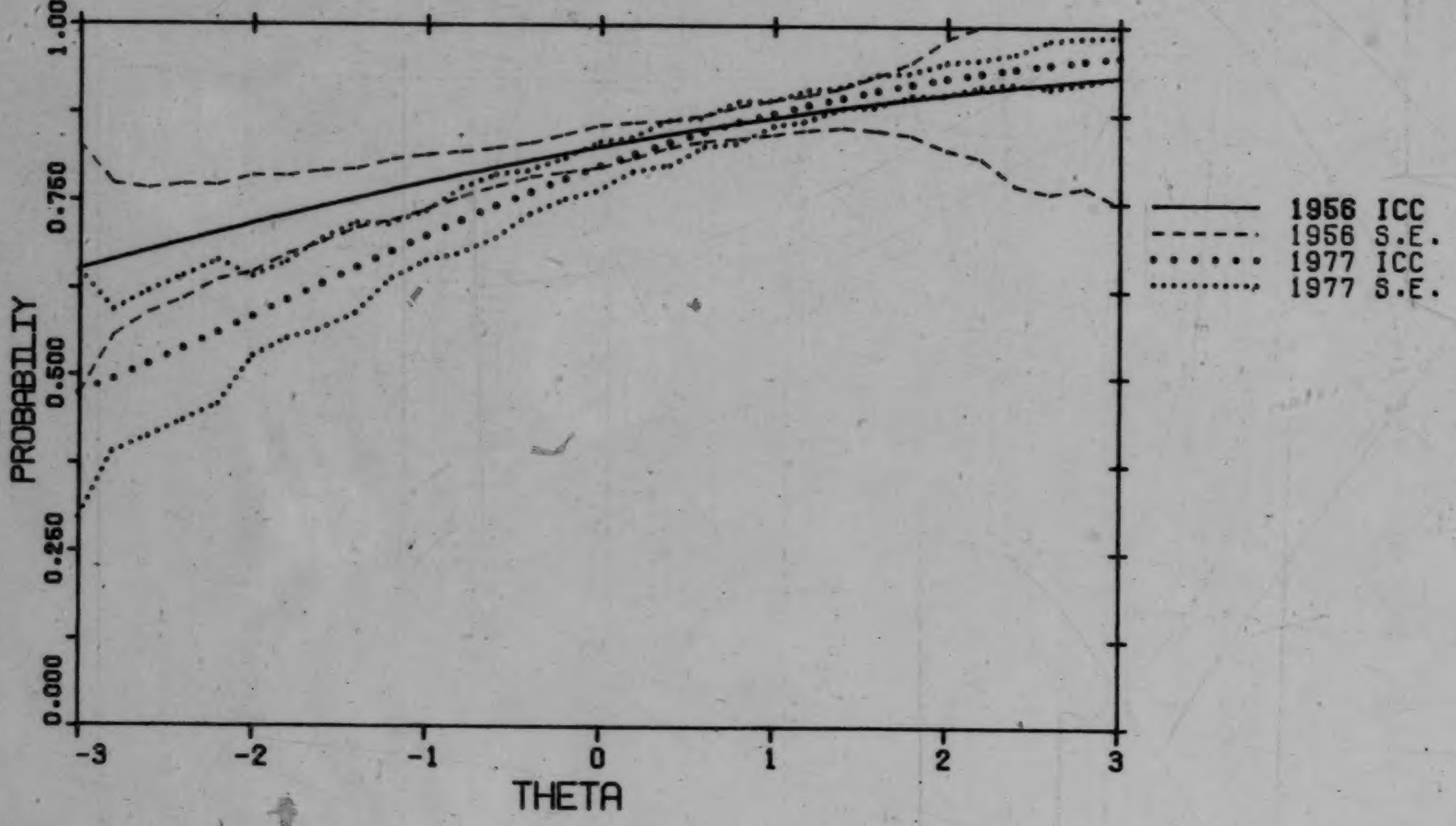


Figure 133



ICC'S FOR ITEM NUMBER 62 : CMM

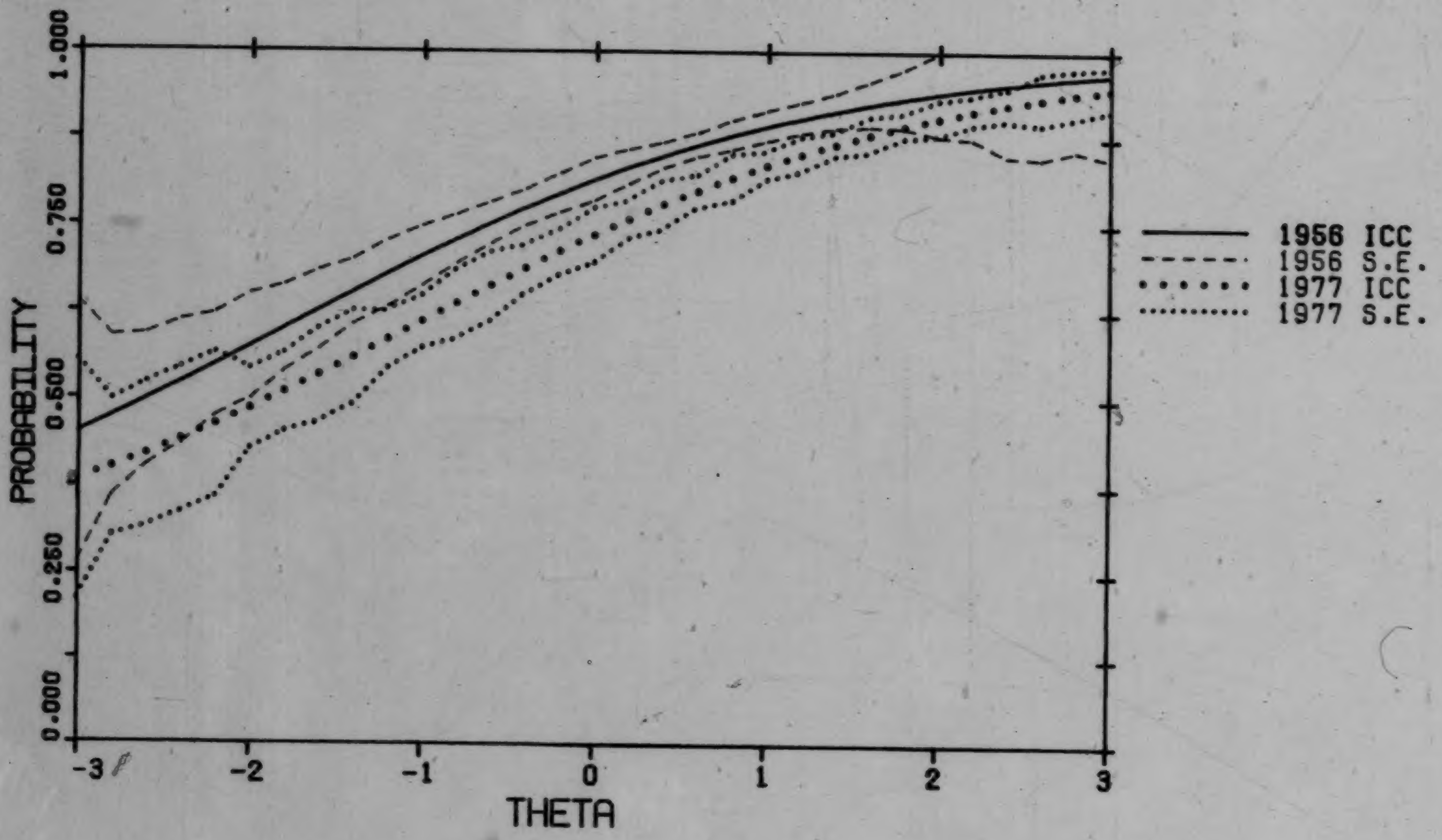


Figure 134

ICC'S FOR ITEM NUMBER 63 : CMM

Figure 135

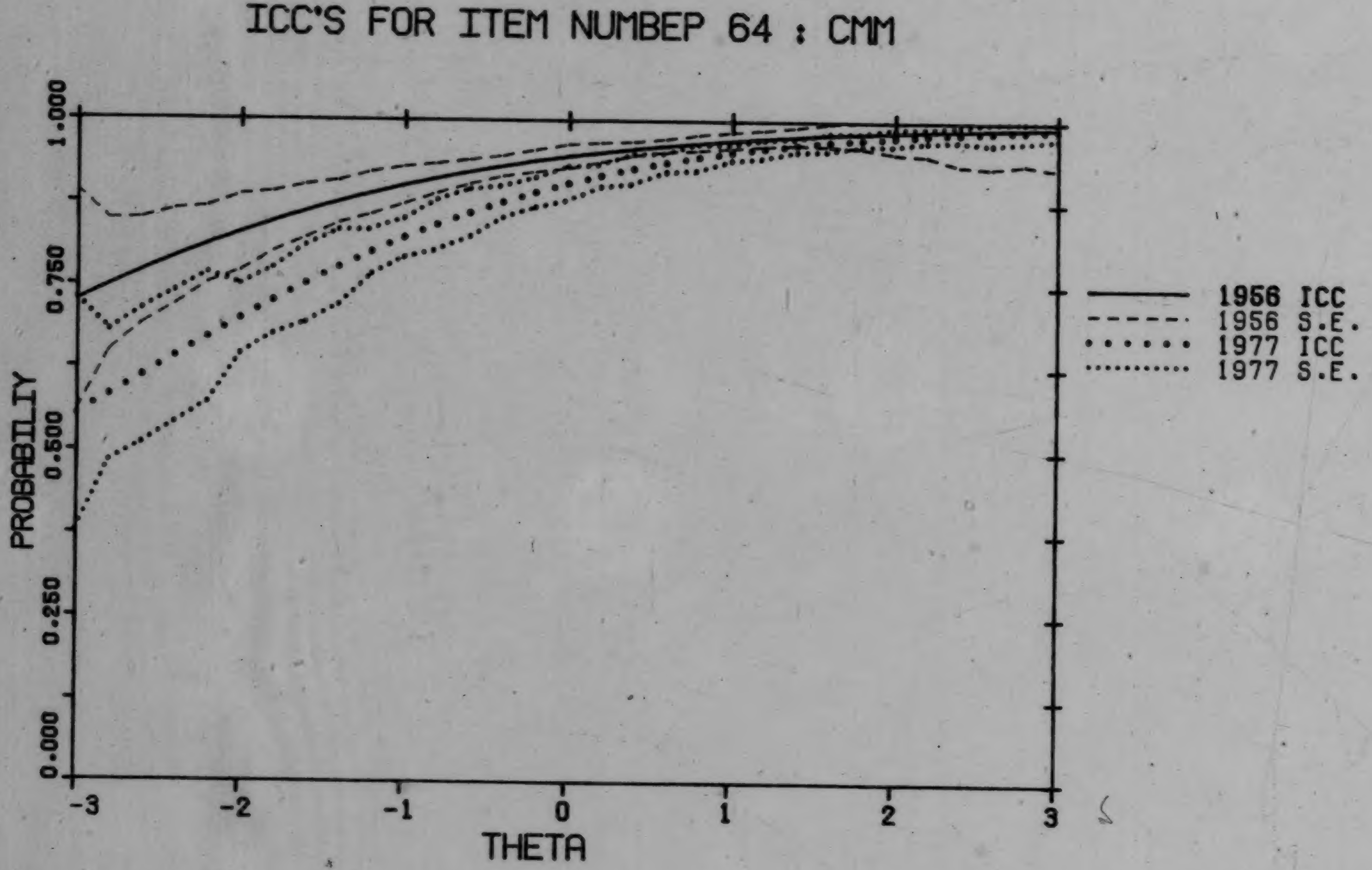

ICC'S FOR ITEM NUMBEP 64 : CMM
PROBABILITY
1.000
0.750
0.500
0.250
0.000
THETA
-3
-2
-1
0
1
2
3
1956 ICC
1956 S.E.
1977 ICC
1977 S.E.

Figure 136

ICC'S FOR ITEM NUMBER 65 : CMM

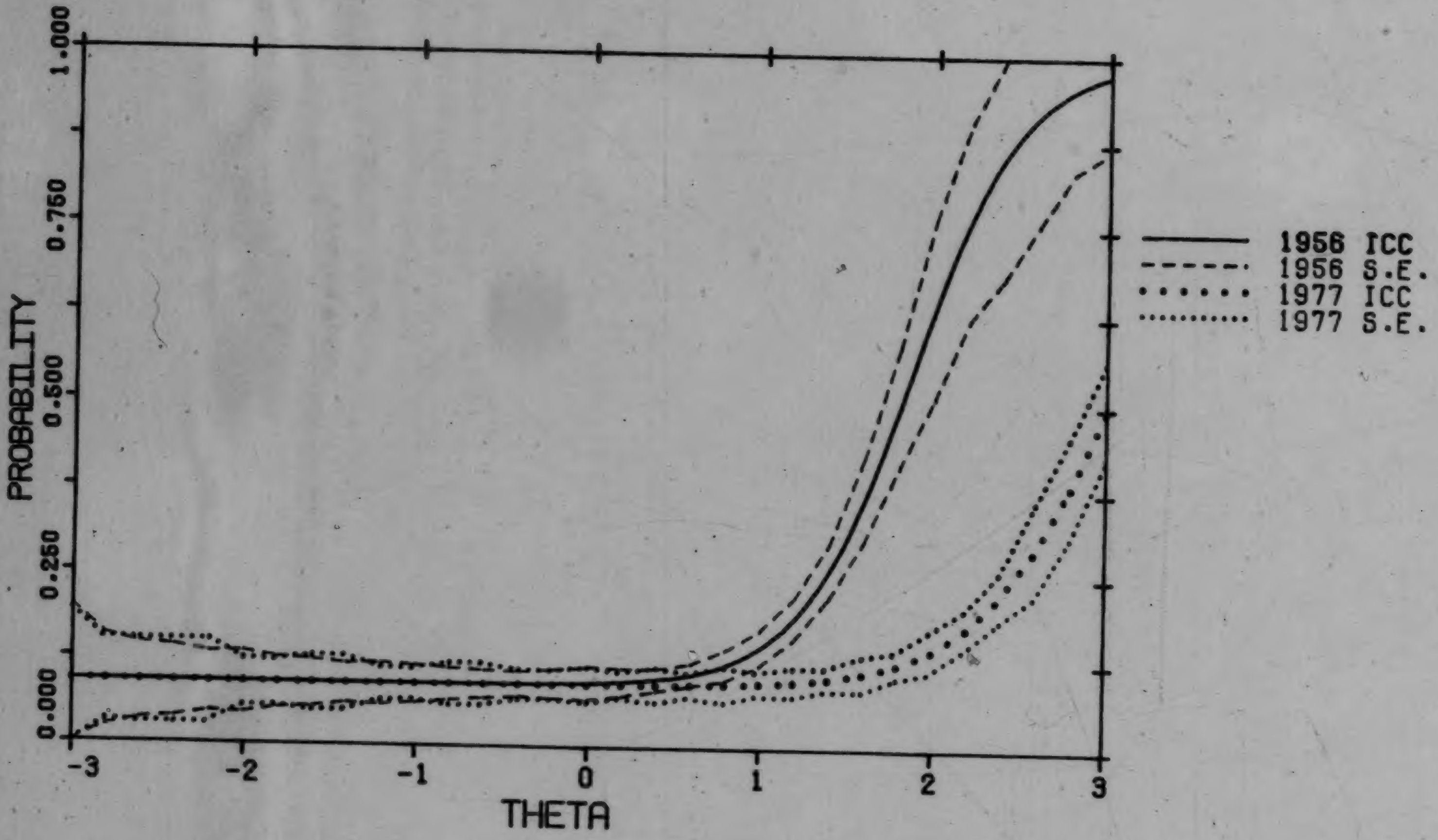


Figure 137

ICC'S FOR ITEM NUMBER 69 : CMM

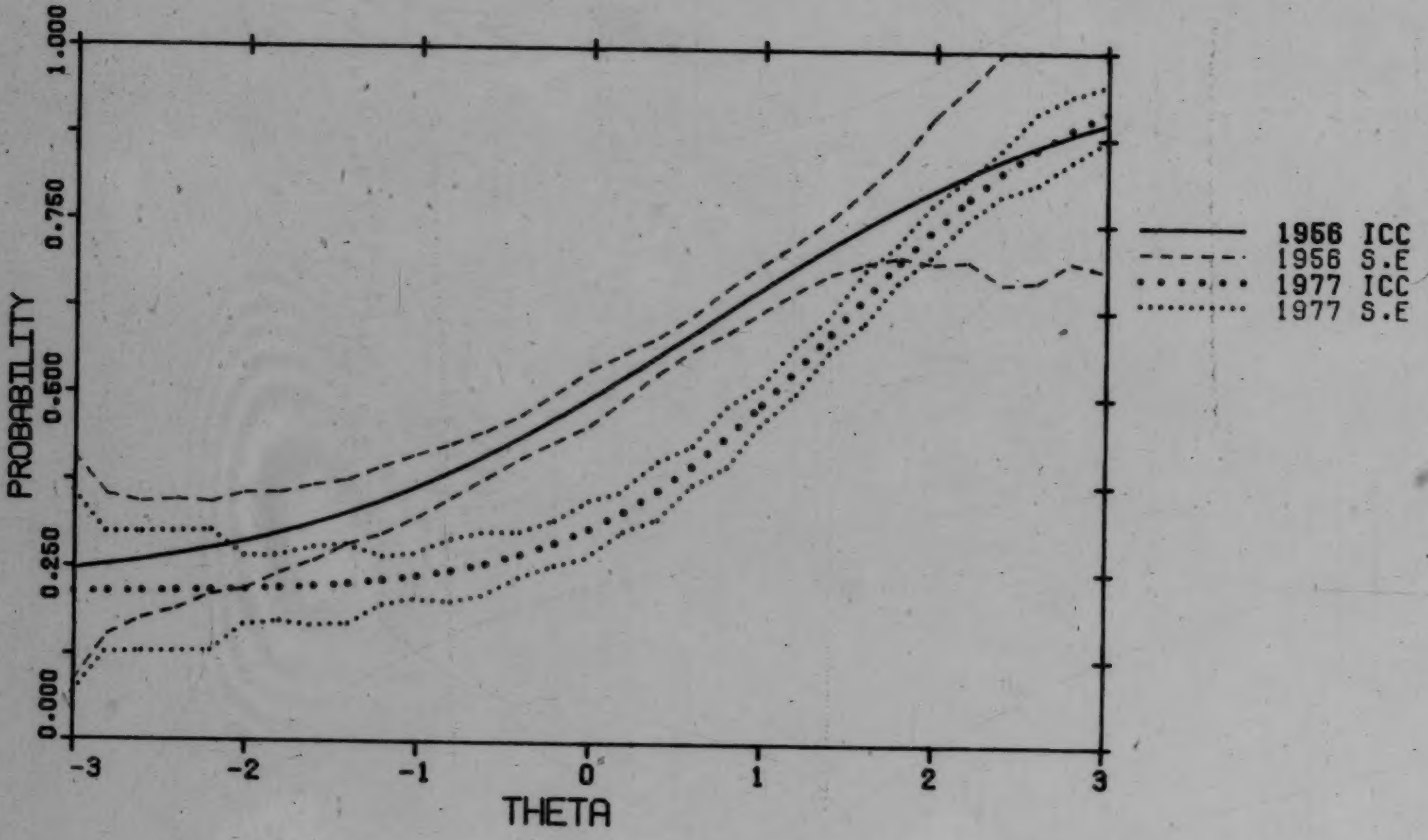


Figure 138

ICC'S FOR ITEM NUMBER 70 : CMM

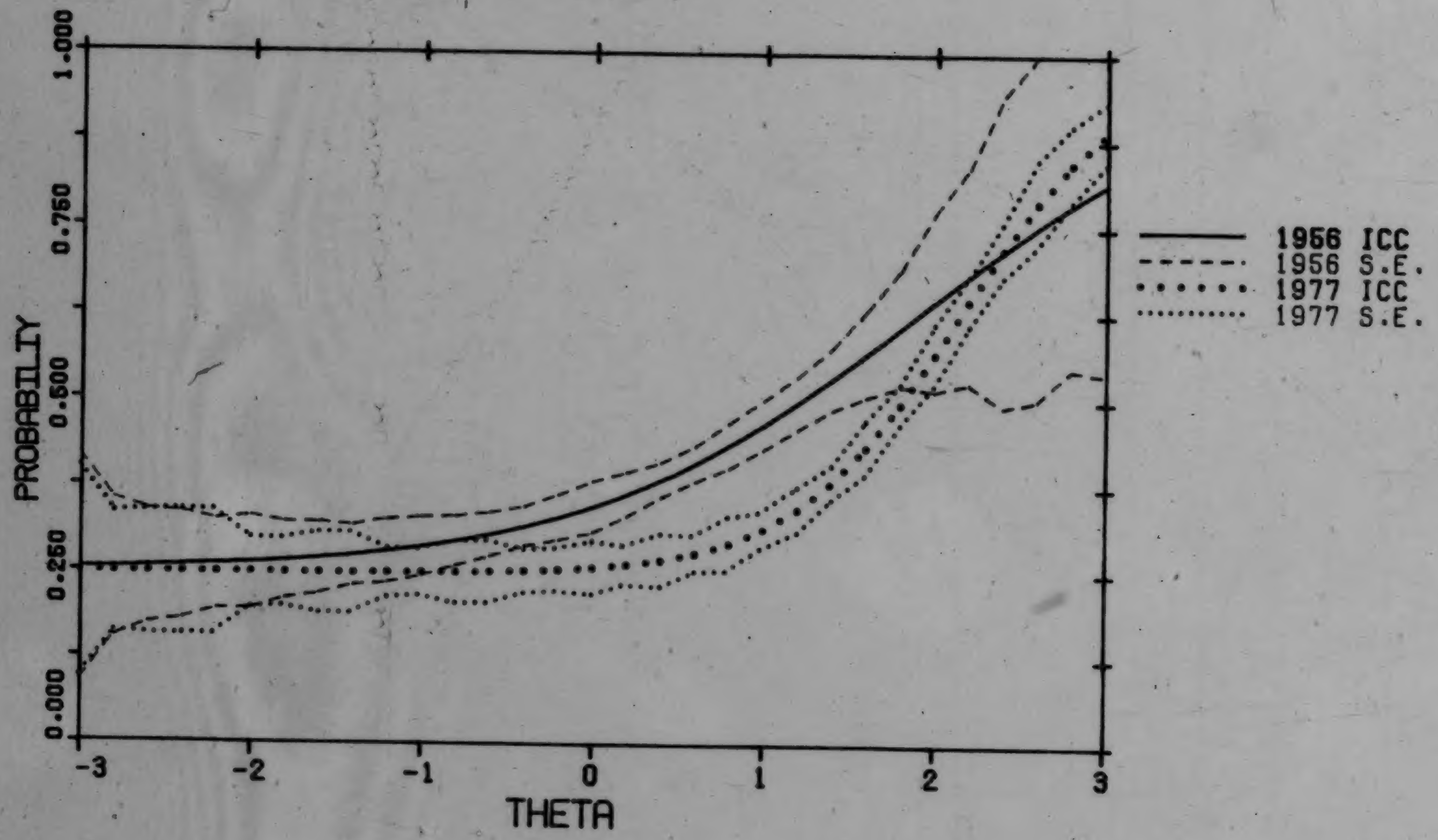


Figure 139

ICC'S FOR ITEM NUMBER 80 : CMM

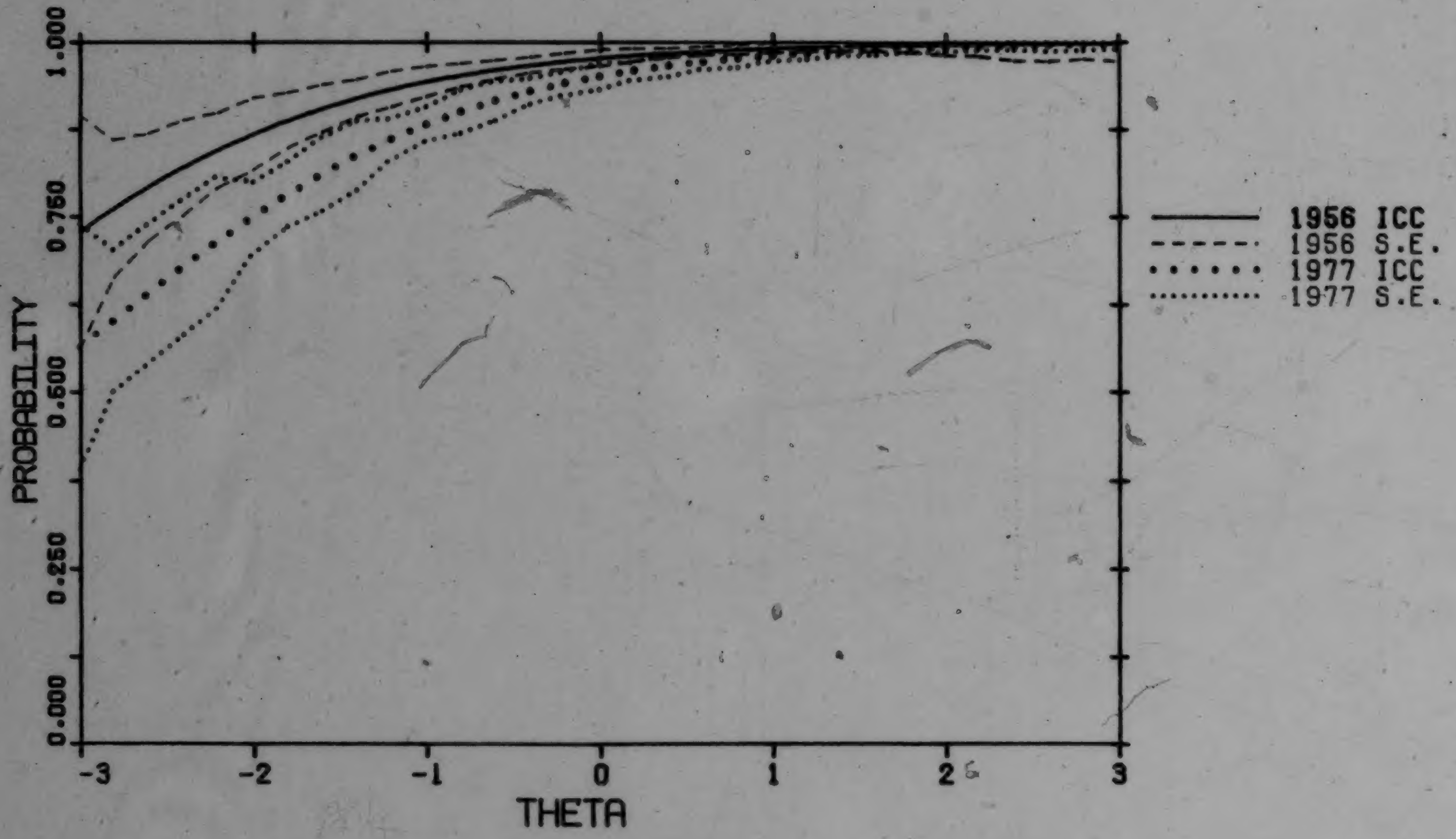


Figure 140

ICC'S FOR ITEM NUMBER 87 : CMM

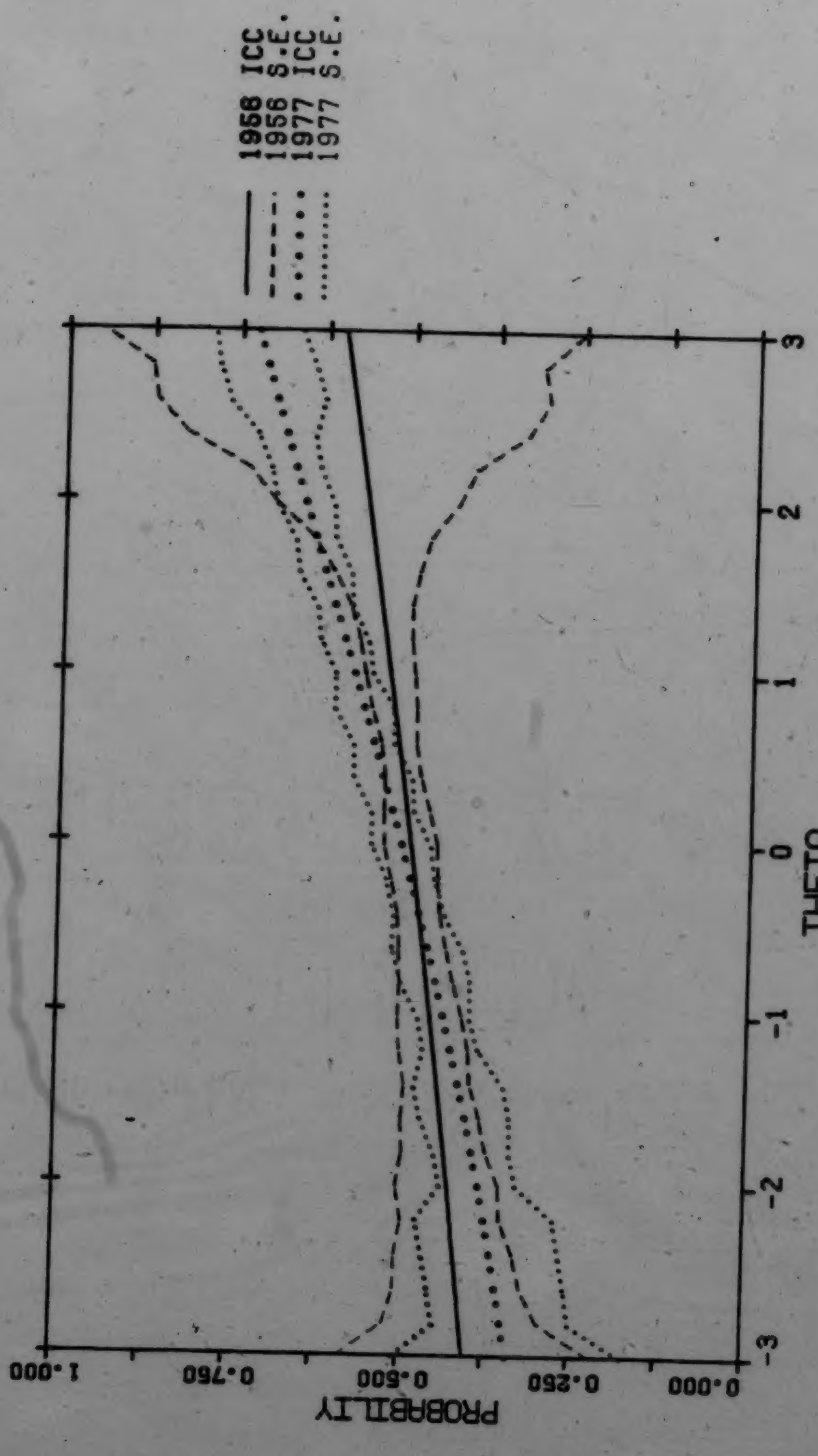


Figure 141

ICC'S FOR ITEM NUMBER 89 : CMM

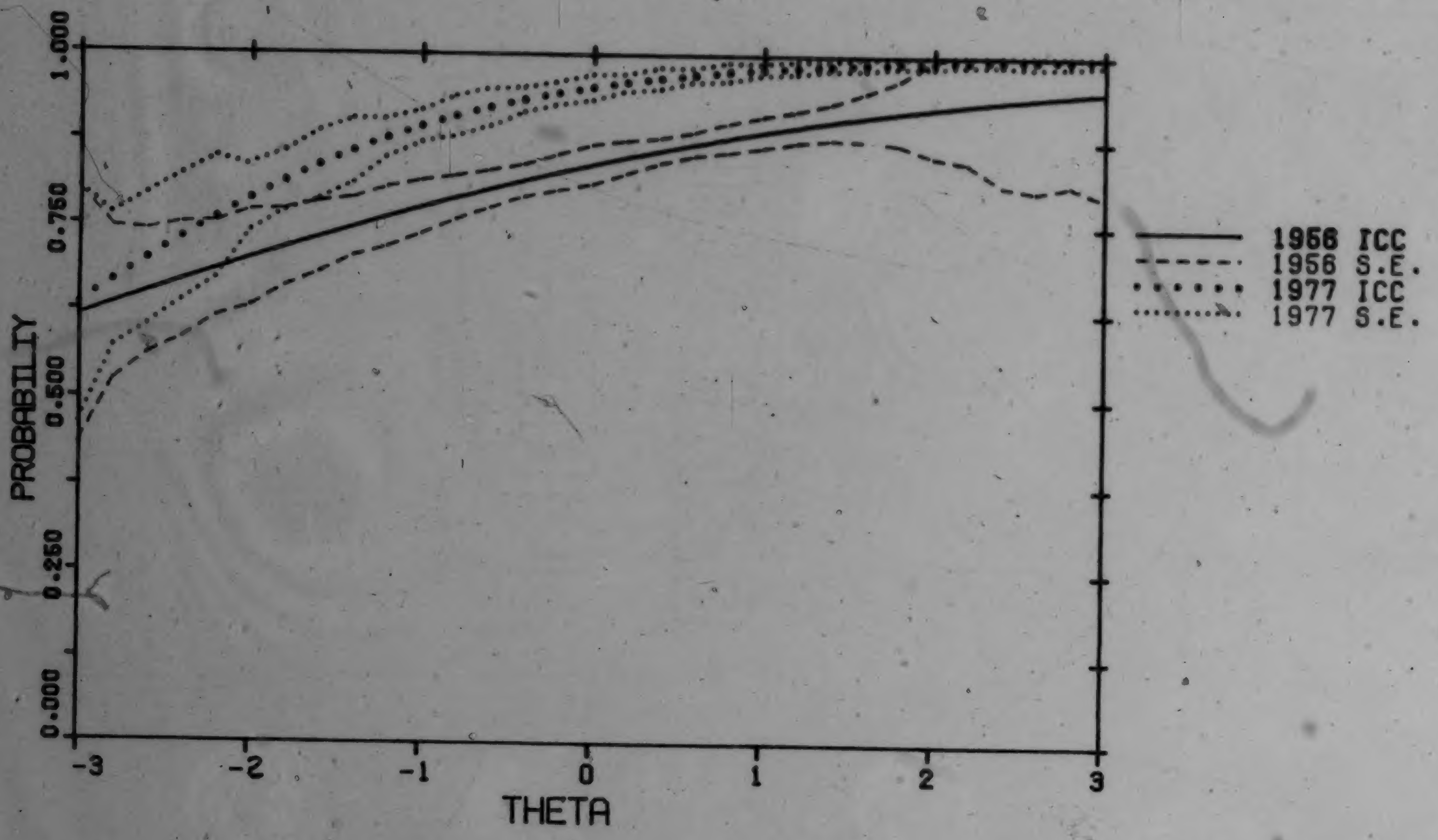


Figure 142

ICC'S FOR ITEM NUMBER 90 : CMM

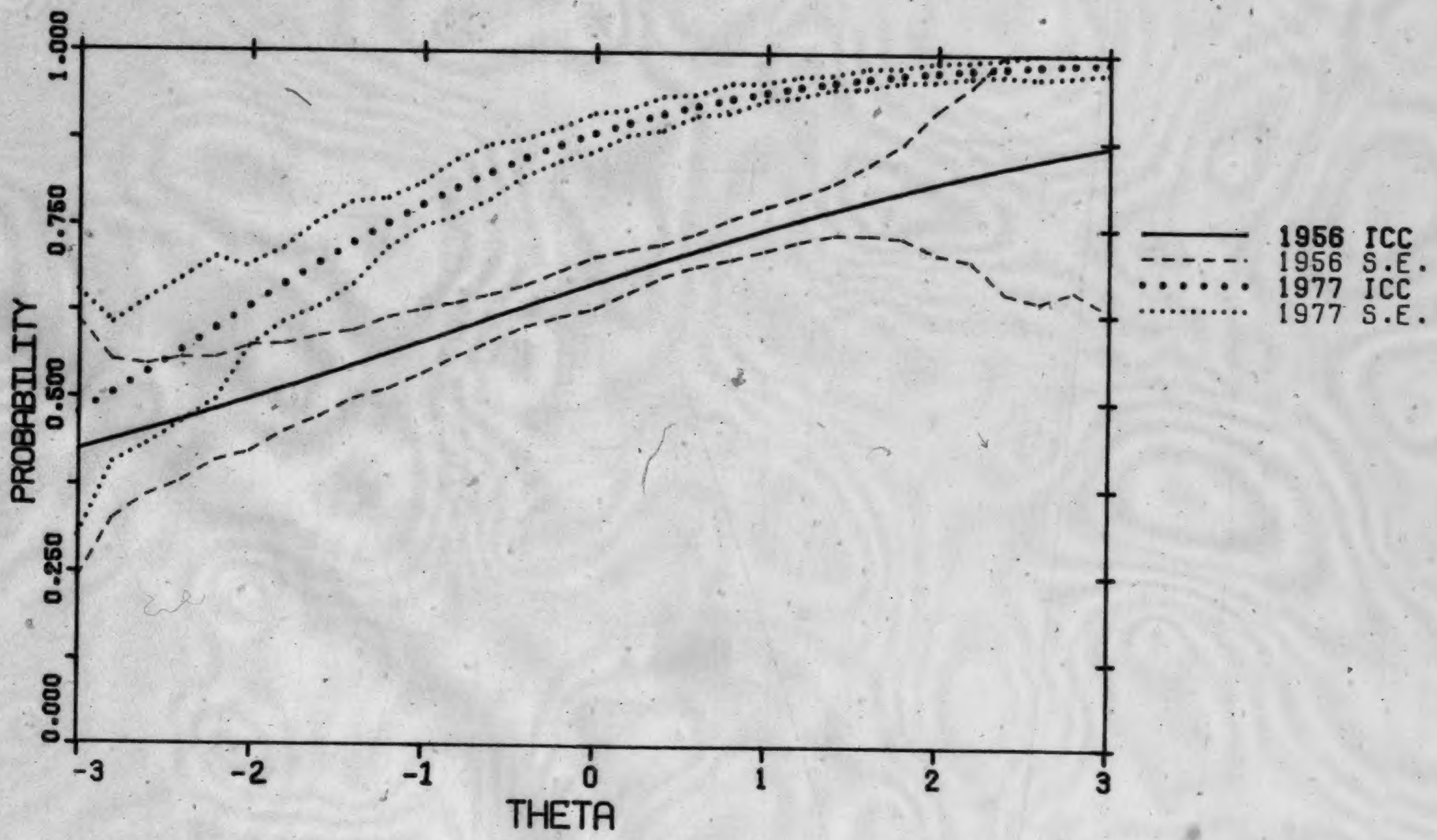


Figure 143

ICC'S FOR ITEM NUMBER 93 : CMM

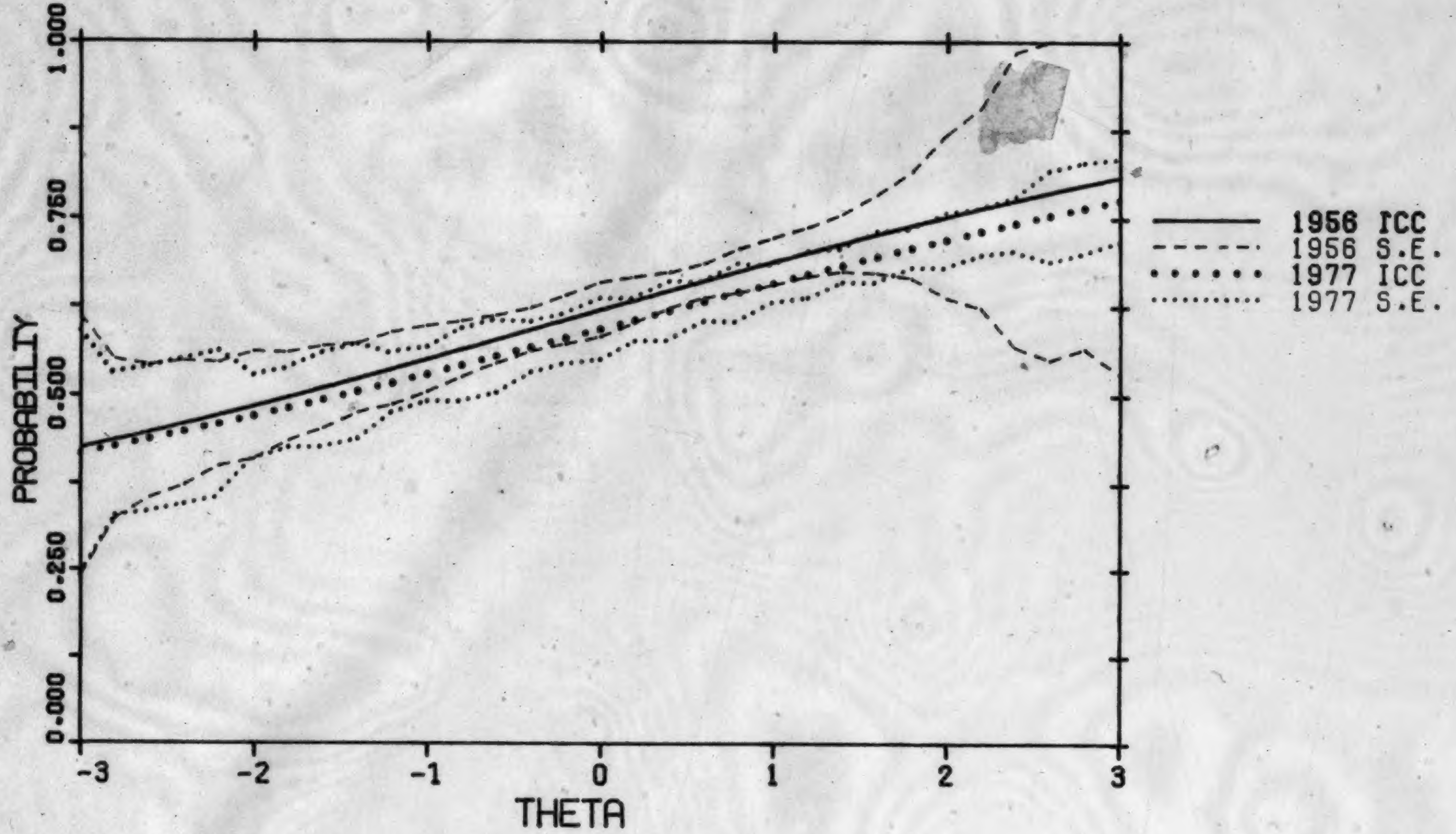


Figure 144

ICC'S FOR ITEM NUMBER 94 : CMM

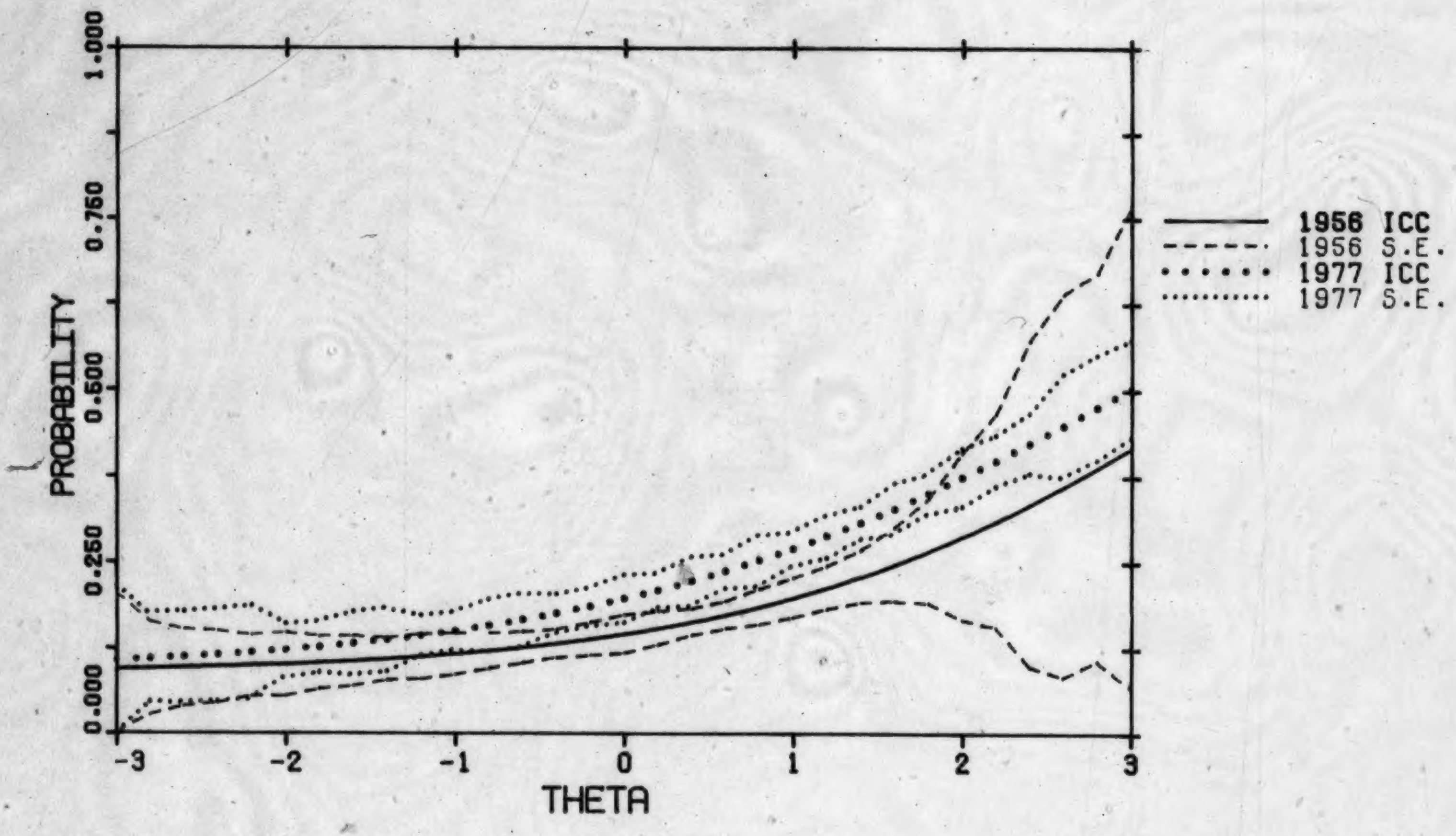


Figure 145

ICC'S FOR ITEM NUMBER 95 : CMM

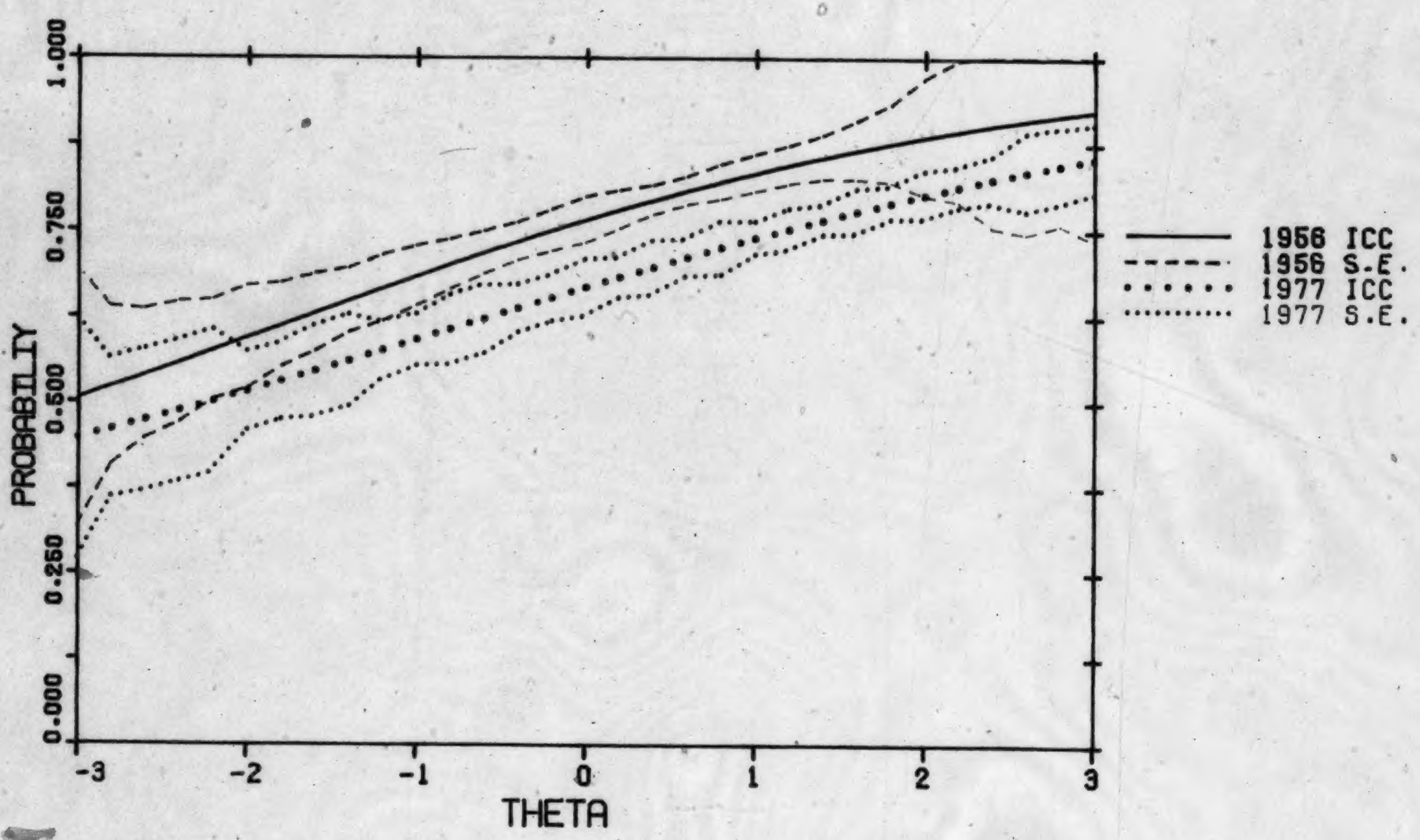


Figure 146

ICC'S FOR ITEM NUMBER 96 : CMM

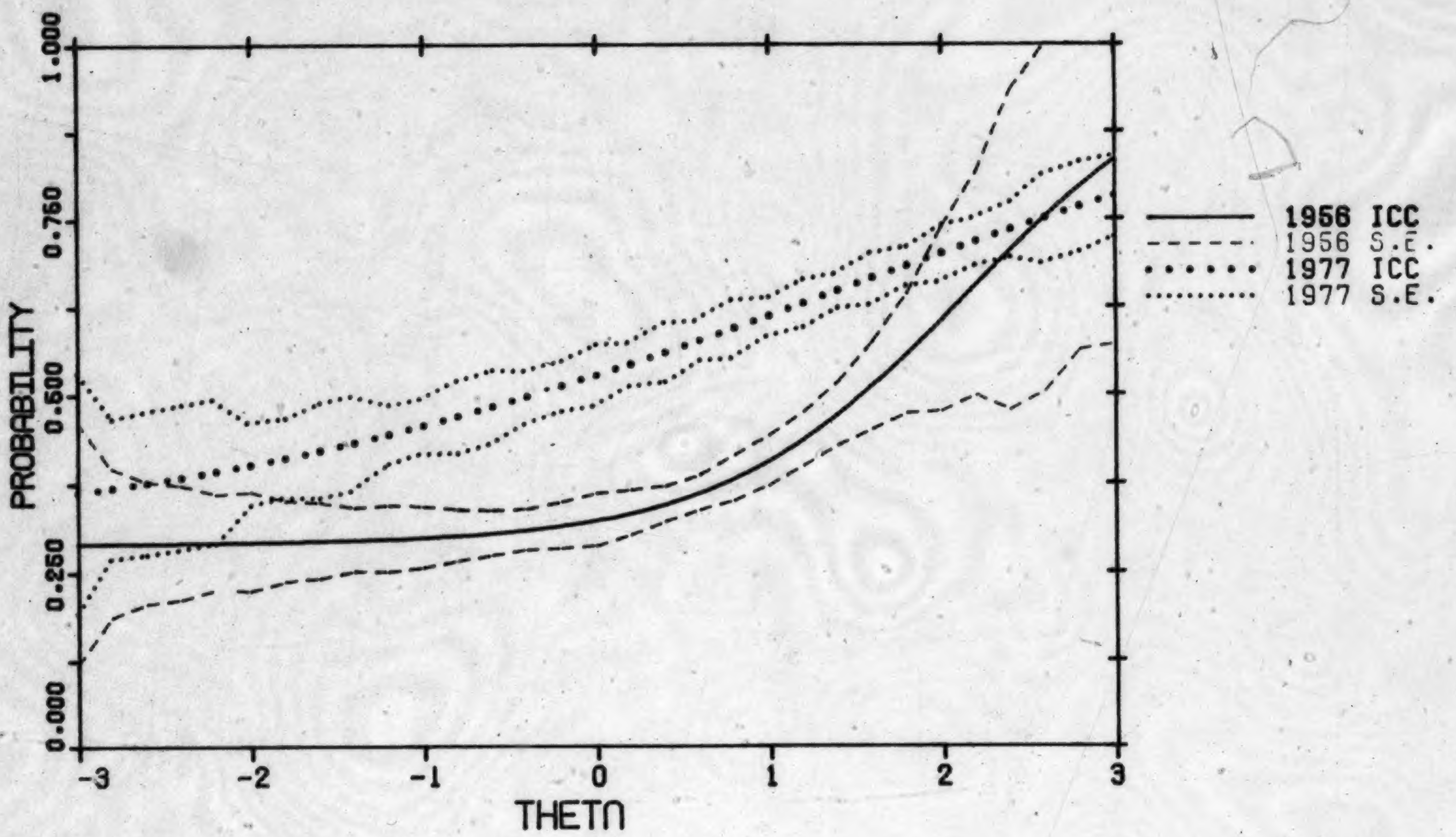


Figure 147

ICC'S FOR ITEM NUMBER 97 : CMM

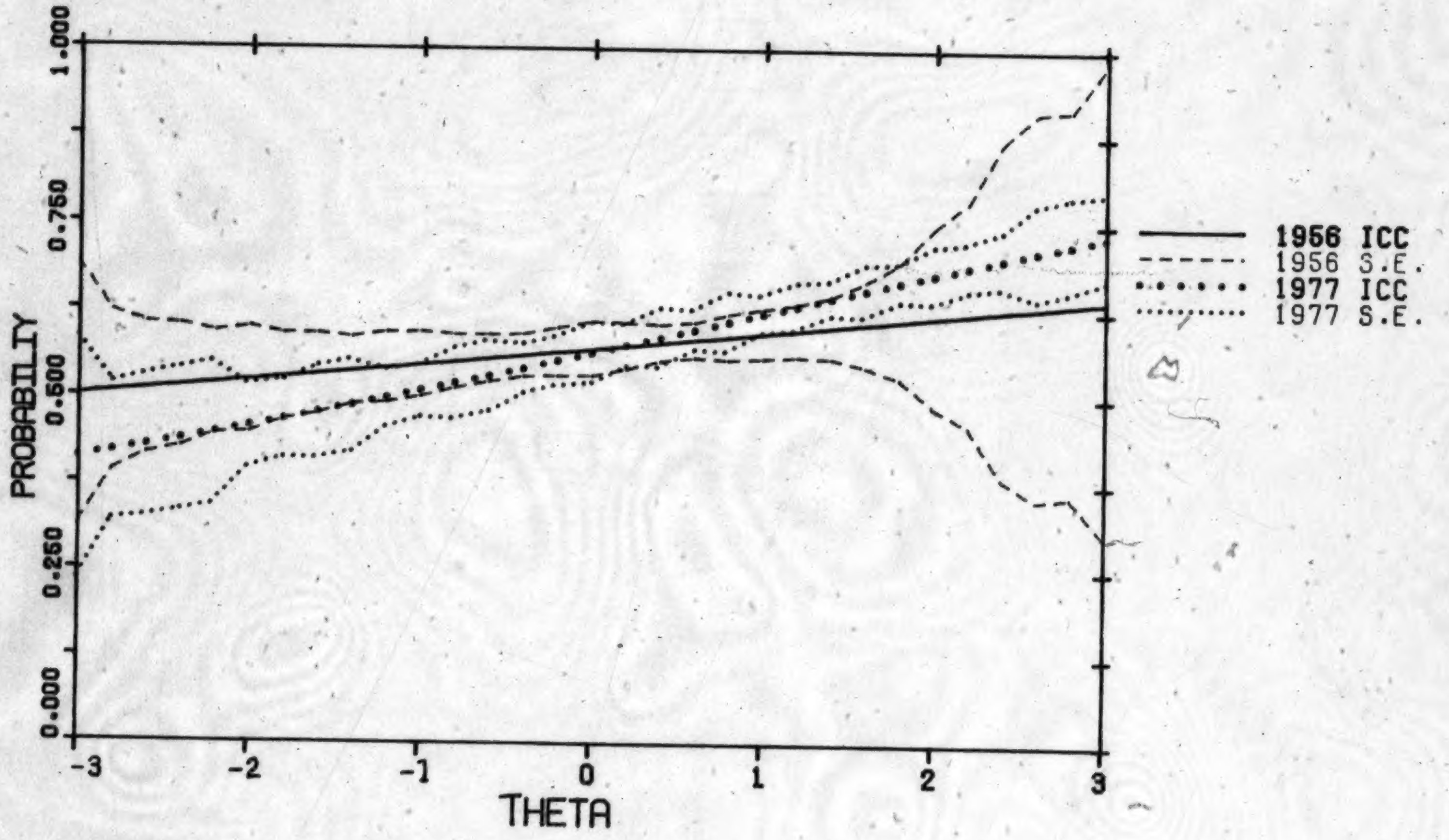
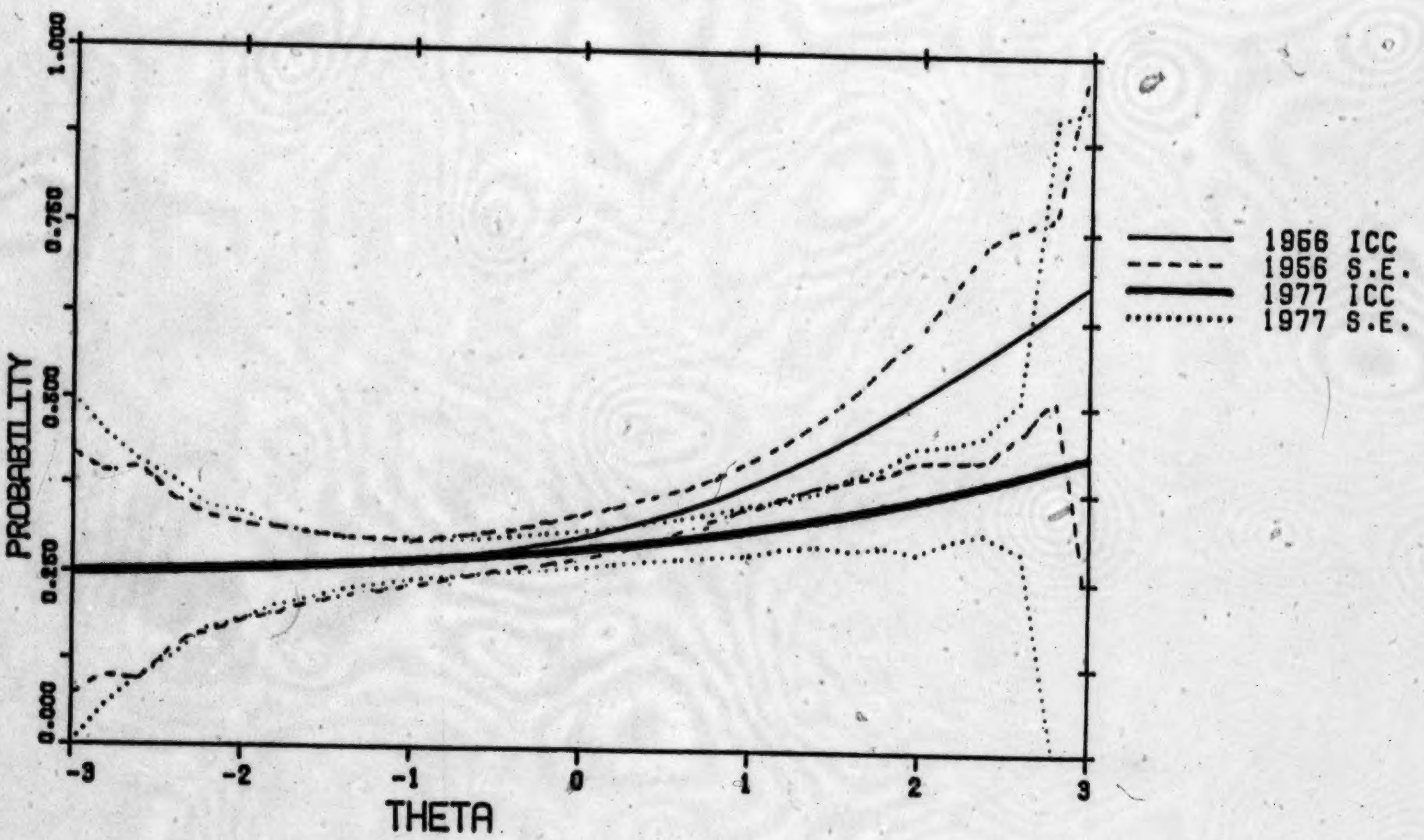


Figure 148

ICC'S FOR ITEM: b ≥ 4.00



References

Allen, M. J., & Yen, W. M. *The introduction to measurement theory*. Monterey, Calif.: Wadsworth, 1979.

Anderson, E. B. Asymptotic properties of conditioned maximum likelihood estimates. *The Journal of the Royal Statistical Society*, Series *B*, 1970, *32*, 283-301.

Angoff, W. Why the S.A.T. scores are going down. *English Journal*, 1975, *64*(3), 10-11.

Birnbaum, A. Some latent trait models and their use in influencing an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item paramaters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29-51.

Bock, R. D., & Liebermann, M. Fitting a response model for n dichotomously scored items. *Psychometrika*, 1970, *35*, 179-197.

Clarke, S. C. T., Nyberg, V., & Worth, W. H. *General report on Edmonton grade III achievement: 1956-1977 comparisons*. A study conducted for the Alberta Advisory Committee on Educational Studies under contract to Alberta Education. Edmonton, Alberta: Alberta Education, 1977.

Clarke, S. C. T., Nyberg, V., & Worth, W. H. *Technical report on Edmonton grade III achievement 1956-1977 comparisons*. A study conducted for the Alberta Advisory Committee on Educational Studies under contract to Alberta Education. Edmonton, Alberta: Alberta Education, 1978.

Gates, A. I. *Gates advanced primary reading tests* for Grade 2 (second half) and Grade 3. Form I. New York: Bureau of publications, Teachers' College, Columbia University, 1942.

Gates, A. I. *Manual of directions: Gates advanced primary reading tests*. New York: Bureau of Publications, Teachers' College, Columbia University, 1943.

Gullikesen, H. *Theory of Mental Tests*. New York: Wiley, 1950.

Hambleton, R. K., & Cook, L. Latent trait models and their use in the analysis of educational data. *Journal of Educational Measurement*, 1977, *14*, 76-96.

Hambelton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. *Developments in latent trait theory: A review of models, technical issues, and applications*. (ERIC Document Reproduction Service No. ED 137 364), 1977.

Hambelton, R. K., & Traub, R. E. Information curves and efficiency of three logistic test modes. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 273-281.

Hambleton, R. K., & Traub, R. E. Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 195-211.

Hedges, H. G. *Achievement in basic skills*. A report funded under contract by the Minister of Education, Ottawa, Ontario, 1977.

Jaeger, R. M. A word about the issue. *Journal of Educational Measurement*, 1977, *14*(2), 73-74.

Lord, F. M. A theory of test scores. *Psychometric Monographs*, 1952, No. 7.

Lord, F. M. The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953, *13*, 517-548.

Lord, F. M. An analysis of the verbal scholastic aptitude tests using Birnbaum's three-parameter logistic mode. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F. M. Power scores estimated by item characteristic curves. *Educational and Psychological Measurement*, 1973, *33*, 219-224.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 1974, *39*, 247-264.

Lord, F. M. *A survey of equating methods based on item characteristic curve theory*. Research Bulletin 75-13. Princeton, N.J.: Educational Testing Service, 1975.

Lord, F. M. A study of item bias, using item characteristic curve theory. (ERIC Document Reproduction Service No. ED 137 486), 1976. (a)

Lord, F. M. A broad-range tailored list of verbal ability. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, D.C.: U.S. Civil Service Commission, 1976. (b)

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14(2), 117-138

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

Mario, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14(2), 139-160.

Muir, Norman, D. A comparison of the competence in albegra of the grade IX students of the Edmonton public schools in 1938 and 1959. Alberta Journal of Educational Research, 1961, 7(4), 175-184.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Raven, J. C. Raven Progressive Matrices, Sets A, Ab, B. London: H. K. Lewis & Co., Ltd., 1947.

Raven, J. C. Guide to using the coloured progressive matrices, Sets A, Ab, B. London: H. K. Lewis & Co., Ltd., 1965.

Rentz, R. R., & Rentz, C. C. Does the Rasch model really work? A discussion for practitioners. Measurement in Education, 1979, 10(2), 1-11.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. (ERIC Document Reproduction Service No. Ed 137 337), 1977.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, 1969, No. 17.

Samejima, F. A general model for free-response data. Psychometric Monograph, 1972, No. 18.

Samejima, F. Homogeneous case of the continuous response model. Psychometrika, 1973, 38, 203-219.

Sullivan, E. T., Clark, W. W., & Tiegs, E. W. California short form test of mental maturity, primary grades 1, 2, 3, 1953 S-Form. Los Angeles: California Test Bureau, 1953.

Tiegs, E. W., & Clark, W. W. California achievement tests complete battery, primary grades 1, 2, 3, and 4. Form CC: Reading, Arithmetic, Language. Los Angeles: California Test Bureau, 1950.

Thissen, D. M. Information in wrong responses to Raven's progressive matrices. Journal of Educational Measurement, 1976, 13, 210-214.

Thurstone, L. L. The reliability and validity of tests. Ann Arbor, Michigan: Edwards Brothers, 1931.

Tyler, R. W. Appraising and recording student progress. In E. R. Smith & R. W. Tyler (Eds.), Adventures in American education, Vol. 3. New York: Harper and Brothers, 1942.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14(2).

Wood, R. L., Wigersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N.J.: Educational Testing Service, 1976.

Wright, B. D. Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. D., & Douglas, G. A. Conditional versus unconditional procedures for sample-free analysis. Educational and Psychological Measurement, 1977.

Wright, B. D., & Mead, R. J. BICAL: Calibrating rating scales with the Rasch model. Research Memorandum No. 23. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B. D., Mead, R., & Draba, R. *Detecting and connecting test item bias with a logistic response mode*. Research Memorandum No. 22. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1976.

APPENDIX A

PREVIOUSLY COPYRIGHTED MATERIAL

IN APPENDIX A, LEAVES 285-299

NOT MICROFILMED.

AVAILABLE FROM:

BUREAU OF PUBLICATIONS, TEACHERS COLLEGE
COLUMBIA UNIVERSITY, NEW YORK
by ARTHUR I. GATES